UNIVERSITY COLLEGE LONDON

Molecular Dynamics Simulations of Nucleotide Translocation through α-Hemolysin Nanopores

by

Hugh S C Martin

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy in Chemistry

in the

Department of Chemistry

February 2011

Declaration of Authorship

I, Hugh S C Martin, declare that this thesis titled, 'Molecular Dynamics Simulations of Nucleotide Translocation through α-Hemolysin Nanopores' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UNIVERSITY COLLEGE LONDON

Abstract

Department of Chemistry

Doctor of Philosophy

by Hugh S C Martin

The translocation of polynucleotides through transmembrane protein pores is a fundamental biological process with important medical and biotechnological relevance. The complex translocation process is influenced by a range of factors including the diameter and inner surface of the pore, the secondary structure of the polymer, and the interactions between the polymer and protein. Computer simulations are an invaluable means to investigate microscopic systems and thereby provide a unique, atomistic perspective of important states and processes. This thesis explores how two molecular dynamics methodologies can simulate the translocation of nucleotides through the nanopore α-hemolysin. In the first methodology, non-equilibrium constant velocity-steered simulations are combined with Jarzynski's identity to derive the free energy profiles for the passage of a polynucleotide molecule through the pore. In the second methodology, the free energy profiles are calculated from a biasing force which varies in response to energy barriers encountered during the simulation. Both approaches are used to explain the experimentally observed differences in translocation time through the nanopore between polyadenosine and polydeoxycytidine. In addition to polynucleotides, the study also investigates single nucleotide translocation. Together, the simulations highlight the role of molecular interactions between the nucleic acid molecules and the protein pore. In particular, we find that specific residues of the protein pore dominate the translocation. The unique data set helps assess two methodologies to simulate a system of considerable size and complexity.

Acknowledgements

First and foremost, I would like to thank my three mentors: Stefan Howorka, Shantenu Jha, and Peter Coveney, who have all been tremendously supportive throughout my PhD. I am extremely fortunate to have had the opportunity to work with such committed and esteemed scientists. I am very grateful to Stefan Howorka for years of relentless support and training in all scientific aspects. I am hugely thankful for Shantenu Jha's constant encouragement, belief, and insight. And my sincerest thanks go to Peter Coveney for the many opportunities he has provided me, and for his understanding during difficult times.

I am indebted to my colleagues, David Wright, Liam McNamara, Tom Hewer, Nick Mitchell, and Helen Kinns for their assistance and advice in various scientific and technical domains, each having aided me in the completion of this degree.

I would like to express my gratitude to my family: Dad, Ma, Tom and Chuck, who have always been there for me. I am particularly indebted to my Father, without his years of support I could not have come this far. Thank you also to Joan Roberts (AJ) for igniting the scientific spark in me.

Thank you to my friends, in addition to those mentioned above, William Brown, Heather Campbell, John Edridge, David Edwards, Alex Leach, Philip Leach, and Chieh Chien Lee, each of whom has helped keep my sanity in check during this degree...

Contents

List of Figures

List of Tables

Abbreviations

αHL	α **H**emo**l**ysin
A	**A**denosine
ABF	**A**daptive **B**iasing **F**orce
C_α	α-**C**arbon
cv-SMD	**c**onstant **v**elocity **S**teered **M**olecular **D**ynamics
dC	**D**eoxy**c**ytidine
DNA	**D**eoxyribo**n**ucleic **A**cid
FEP	**F**ree **E**nergy **P**erturbation
JE	**J**arzynski's **E**quality
MD	**M**olecular **D**ynamics
NAMD	**Na**noscale **M**olecular **D**ynamics
PME	**P**article-**m**esh **E**wald
PMF	**P**otential of **M**ean **F**orce
RMSD	**R**oot **M**ean **S**quare **D**eviation
RNA	**R**ibo**n**ucleic **A**cid
SCCR	**S**ingle **C**hannel **C**urrent **R**recording
SMD	**S**teered **M**olecular **D**ynamics
US	**U**mbrealla **S**ampling
VMD	**V**isual **M**olecular **D**ynamics

Symbols

A	Helmholtz Free Energy	J mol^{-1}
f	Force	N
F	Free Energy	J mol^{-1}
G	Gibbs Free Energy	J mol^{-1}
J	Jacobian	
k_B	Boltzmann Constant	J K^{-1}
p	Pressure	Pa
P	Probability	
Q	Partition Function	
S	Entropy	J K^{-1}
SE	Standard Error	
T	Temperature	K
V	Volume	m^3
W	Work	J

$\mathcal{H}$	Hamiltonian	
$\mathcal{P}$	Probability Density	
$\mathcal{V}$	Potential energy	J
β	Inverse Temperature	K J^{-1}
λ	Order Parameter	
ξ	Reaction Coordinate	
σ	Standard Deviation	
ζ	Instantaneous Force Measurement	
	Threshold Parameter	

Dedicated to my beloved soulmate, Hinnah...

Chapter 1

Introduction

1.1 Polynucleotides

Polynucleotides in nature exist in two classes: deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) [1]. DNA stores genetic information for the development and function of most living organisms and some viruses. RNA is transcribed from DNA by enzymes in the production of proteins, and most viruses keep their genetic information in the form of RNA.

The molecular structure of DNA is shown on the left of Figure 1.1. Each nucleotide monomer in the polymeric chain is composed of a nucleotide base, a pentose sugar (deoxyribose), and a phosphate group [1]. The deoxyribose sugar and phosphate group form the repeat unit of the polynucleotide backbone, with each monomer connecting to the next via a phosphodiester bond. At physiological pH, the phosphate groups in the polynucleotide chain possess a single negative charge while terminal phosphate groups posses a double negative charge. The terminus that ends with a phosphate group is called the $5'$-end due to the relation to the pentose sugar. The opposite end, which is terminated at a pentose sugar, is termed the $3'$-end. Attached to the pentose sugar is one of four types of nucleotide base: cytosine, guanine, adenine and thymine. These nucleotide bases can form hydrogen bonds to complementary bases. This interaction is specific due to the matching molecular structures and hydrogen bonding pattern between guanine and cytosine, or adenine and thymine.

DNA is usually in the double helical form as illustrated on the right of Figure 1.1. In this form, one polynucleotide strand is bound to a second strand of complementary sequence [1]. The shape of the double helix reflects the conformational preferences of the nucleotides as well as the intramolecular interactions between nucleotides within one single strand and intermolecular interactions between the two complementary strands. A dominating factor which maintains the intermolecular interaction is the energetically favourable base stacking between adjacent bases. Base stacking is determined by three factors: attractive London dispersion forces, short-range exchange repulsion (which is reduced by increased π-orbital overlap), and electrostatic interactions [2]. The sequence of nucleotide bases along the polynucleotide strand encodes the genetic information to form polypeptides.

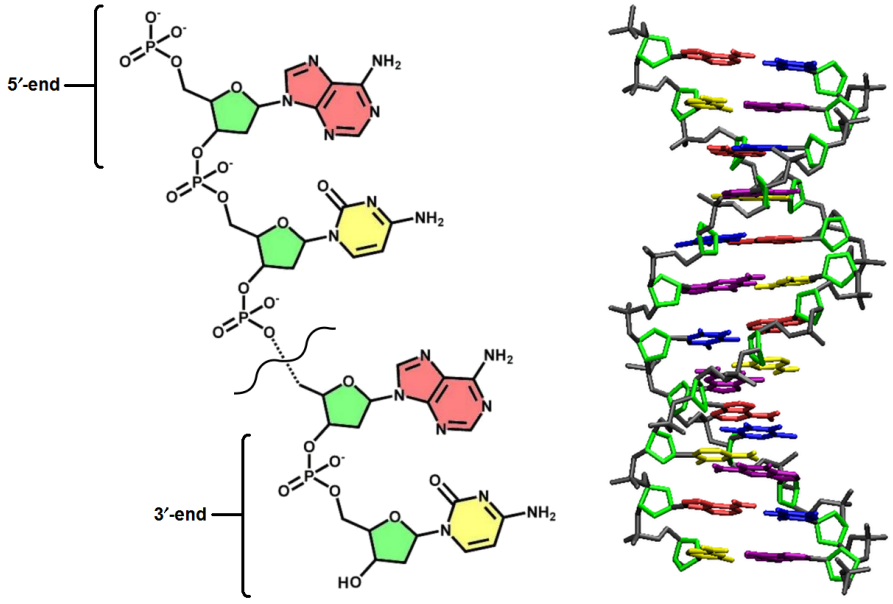


FIGURE 1.1: Illustration of the molecular structure and conformation of DNA. The molecular structure of a single strand of DNA is shown on the left, featuring the phosphate groups, the deoxyribose sugar (green), adenine nucleotide bases (light red) and the cytosine nucleotide bases (yellow). The 3′ and 5′-ends of the polynucleotides are also labelled for reference. The generic double stranded DNA helix conformation is shown on the right. Image (right) prepared using VMD [3].

In the presence of the necessary conditions and molecular machinery, DNA is transcribed into RNA, which in turn serves as a template for the biosynthesis of polypeptides [1]. Similar to DNA, each RNA nucleotide is made up of a pentose sugar, a phosphate, and a base. Unlike DNA, the pentose sugar of RNA is ribose and contains a hydroxyl group at the 2′-carbon of the pentose. Furthermore, the uracil base is used in place of thymine. RNA does not typically exist as a duplex, as found in DNA. However, single stranded

RNA can form secondary structure elements such as hairpin loops, stems, bulges, and internal loops, which are largely mediated via hydrogen bond base pairing.

In single stranded sections of RNA, the biopolymer's conformation is influenced by the energetically favourable overlap of the aromatic π-orbitals on the nucleotide bases. As this interaction is considerably weaker than the hydrogen bonds within a double-stranded structure, single stranded polynucleotides do not tend to form highly regular configurations. Nevertheless, the base-stacking area between adenines is one of the reasons why polynucleotides composed purely of adenosine adopt a helical shape; the hydroxyl of the ribose unit also contributes to the structure. By contrast, the smaller cytosine base in polydeoxycytidine tends to adopt a more random configuration [4–7].

1.2 Proteins

Proteins are folded biopolymers composed of interlinked amino acids [8]. Proteins are key to the structure and composition of all living organisms and participate in almost every biological function. Amino acids possess amine and carboxyl functional groups, which form the amide bond between separate amino acids of the polypeptide chain. Amino acids also contain a side chain which varies between different amino acids types, side chains for lysine and methionine are shown as examples in Figure 1.2. Twenty different amino acids occur in nature, each pertaining different properties depending on the type of side chain attached. Relevant to the work in this thesis, the side chains can be neutral or charged depending on the pH of solution that the amino acid is in.

The linear sequence of amino acids in a polypeptide chain represents the primary structure of proteins [8]. The polypeptide chain can form secondary structures such as α-helices (arising from hydrogen bonds between the amine and carbonyl group of nearby amide bonds) and β-sheets (sections of polypeptide connected laterally by more than 4 hydrogen bonds). The polypeptide, secondary structure included, can fold in itself forming the tertiary structure of the protein. Several proteins may bind together to form a quaternary structure. For instance, the lipid bilayer membrane protein α-hemolysin (αHL) is formed by seven polypeptide chains within a heptameric structure.

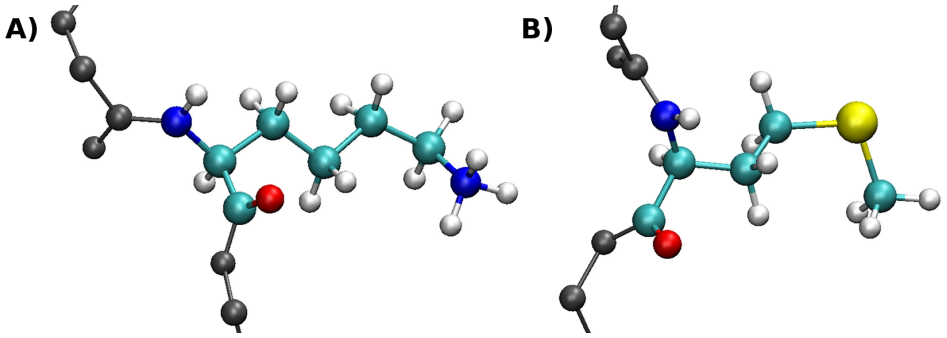


FIGURE 1.2: Representation of the molecular structure of amino acids lysine and methionine. Amino acids are composed of amine and carboxyl functional groups along a polypeptide backbone (grey). They also posses side-chains, the composition of which varies between types of amino acid. An amino acid contains carbon (cyan), nitrogen (blue), and hydrogen (white), and may contain sulphur (yellow). A) Lysine is a positively charged amino acid at physiological pH, due to the terminal amino group of the side chain being protonated. B) Methionine amino acid is a neutral amino acid at physiological pH. Image prepared using VMD [3].

Protein structures can be determined using X-ray crystallography or nuclear magnetic resonance spectroscopy among other methods [9]. By obtaining the atomic coordinates for the atoms in a protein, the protein is then able to be used in computer simulations.

1.3 Polynucleotide Translocation through Nanopores: Biology and Sequencing Application

Nucleic acids and proteins interact to form biologically relevant complexes. Such complexes include those formed during the enzymatic polymerisation of DNA strands or when mRNA is synthesised by proteinaceous polymerases. Another important transient complex involves the passage of DNA strands through a membrane protein. This translocation of polynucleotides across transmembrane pores occurs in several microbiological processes such as phage infection and bacterial conjugation [10, 11]. An example of phage injection is the Enterobacteria phage T4, which is known to infect *E. coli* bacteria by attaching itself to the outer membrane of a cell; forming a membrane channel through which it transfers DNA [10]. In the case of bacterial conjugation, genetic material is transferred between bacterial cells by a bridge-like protein-based connection between two cells [10]. Other examples of polynucleotide transfer across membranes in nature include the uptake of polynucleotides by the kidney, liver and spleen [12] and the transport of polynucleotides across nucleoporins in all eukaryotic cells [13, 14]. It would be

valuable to further understand these naturally occurring translocation processes, with particular relevance to gene therapy [15].

The translocation of DNA through a protein pore is also of interest for a new genetic sequencing approach. In the proposed sequencing strategy, individual DNA strands are pulled through a small nanoscale pore to sequentially read the passing bases. As discussed in greater detail later, the nanopore-based approach has several advantages such as the ability to read native DNA which could help avoid the costly labelling and amplification of DNA in conventional sequencing approaches. In general, there is a great interest in next-generation sequencing methods which can help make the reading of genomes cheaper and more widespread.

Genomes are the blueprint for life and the key for understanding the evolutionary past of the living world [16]. The human genome is about three billion base pairs long and is encoded in DNA. If the genome sequence were written on A4 paper it would be about 400,000 pages long. Small variations between personal genomes lie at the root of a person's individuality. Furthermore, the genetic make-up of a person also contains critical information about the formation and susceptibility of diseases and in the response to chemicals such as drugs [16].

It is apparent that the future of medicine will be strongly influenced by the development of genomics [17–20]. The advancement of genomics will allow new diagnostic methods and the identification of drug targets. For example, a publication in 2007 reported on the tagging of more than 20 genetic markers associated with common diseases by the analysis of 17,000 human DNA samples [21]. Approximately 100,000 somatic mutations from cancer genomes have been reported over the last 25 years, in the coming few years, hundreds of millions more are expected to be revealed through cancer genome sequencing [22]. While the hereditary information is largely based in the DNA sequence of a genome, important traits can also be influenced by epigenetics and gene regulation. Proteomics and post-translational modification are equally important to understand the biological effects of genes.

The Human Genome Project was initially performed using the Sanger chain termination method [23]. Over the course of the 10 year project, a first-draft was obtained at a cost $3 billion [24]. At the time of writing there are 1196 published complete genomes (ranging from bacterial pathogen to human) [25]. With the technology available in

2006, the sequencing of a human genome cost about $20 million [26]. Currently, there is sequencing technology with the potential to sequence genomes at $4400 each [27]. To realise routine medical genome sequencing, a cost target of $1000 per genome by 2014 was set by the National Institutes of Health [28]. With this target now fast approaching, senior scientists such as Richard Lifton and Dr. Rory Collins have begun postulating about $100 genome sequencing [29, 30] .

Progress towards faster and cheaper ensemble sequencing techniques is constantly being made [26]. Potential alternatives to Sanger chain termination sequencing (or variations of it) include fluorescence detection [31, 32] in combination with nanostructures. For example, Eid *et al.* have used zero-mode waveguide nanostructure arrays containing DNA polymerase to perform real time polynucleotide sequencing at the single molecule level with an average accuracy of 99.3% [32]. In this technique, the polymerase molecules performed template-directed synthesis using fluorescently labelled deoxyribonucleoside triphosphates, providing sequence specific fluorescent signals which were detected in real time. Furthermore, Drmanac *et al.* have used combinatorial probe anchor ligation chemistry to assay bases from patterned nanoarrays of self-assembling DNA nanoballs [27]. This sequencing technique produces 1 false reading per 100,000 bases and estimates a cost of $4400 to sequence a human genome, which may be achievable within a single day.

An alternative technique that does not use DNA polymerisation with fluorescence-labelled nucleotides is nanopore recordings [33, 34]. While nanopore recordings are not yet capable of sequencing polynucleotides, it is a fast developing field that represents a potentially important form of genetic sequencing. This technique measures the ionic current flowing through a single pore and is explained in greater detail in the subsequent section.

1.4 Single Channel Current Recordings

Measuring ionic current passing through a single ion channel has been routinely used in neurobiology and biophysics, dating back to 1970 for the purposes of characterising ion channels that play a key role in neuronal signal transduction [35, 36]. Rather than using current recordings as a tool to examine protein pores, Kasianowicz *et al.* in 1996

proposed an alteration to the methodology in order to gather data on the composition of matter passing through the pore. In particular, it was suggested that current recordings could investigate nucleic acid strands that pass through a pore [37]. This approach is illustrated in Figure 1.3. A single nanopore is embedded in a lipid bilayer membrane that separates two electrolyte-filled Teflon reservoirs. The application of a transmembrane potential leads to the translocation of ions and small molecules, such as nucleic acids, through the pore. The translocating molecules can block the movement of ions through the pore, reducing the measurable current flow. By outputting the current flow in the form of a current trace, such blockage events can be detected. Examples of current traces in this manner are shown in Figure 1.4. The degree to which the current flow is reduced, and the length of the blockage event, is dependant on the nature of the translocating molecule. Therefore, the current trace can be used to gain information on the composition of the translocating molecule.

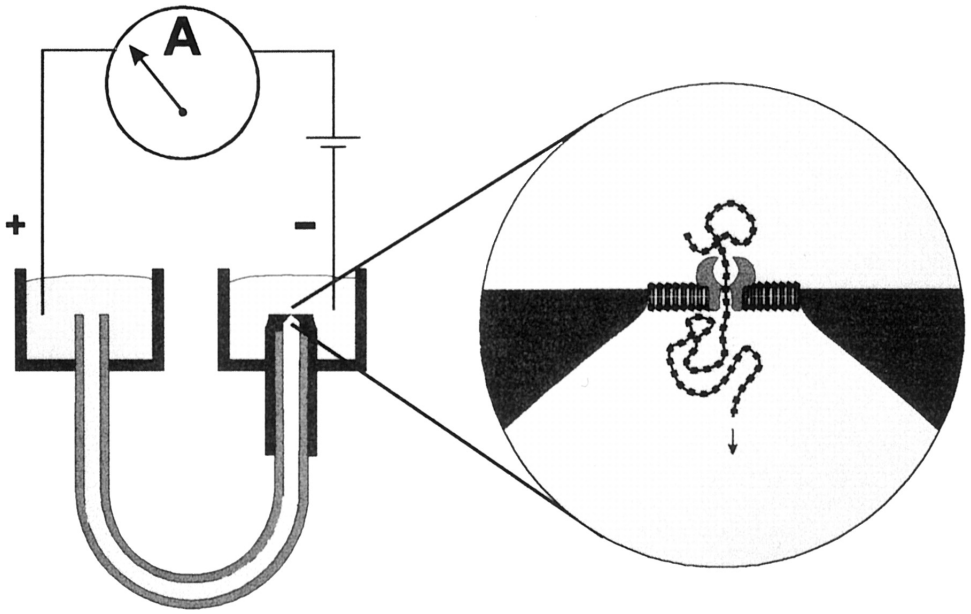


FIGURE 1.3: Representation of a single channel current recording apparatus [37]. Two Teflon reservoirs are separated by a lipid bilayer. A nanopore, which is inserted into the lipid bilayer, is the only conduit through which ionic current can flow between the reservoirs. By applying a transmembrane potential and measuring the ionic current flow between the two reservoirs, it is possible to retrieve information on molecules inside the nanopore. The negative labelled reservoir, was designated the *cis*-side.

The ground breaking piece of work by Kasianowicz *et al.* demonstrated that the duration of the current trace blockage events in SCCR experiments was proportional to the length of the translocating polymer. It follows that with improved experimental components

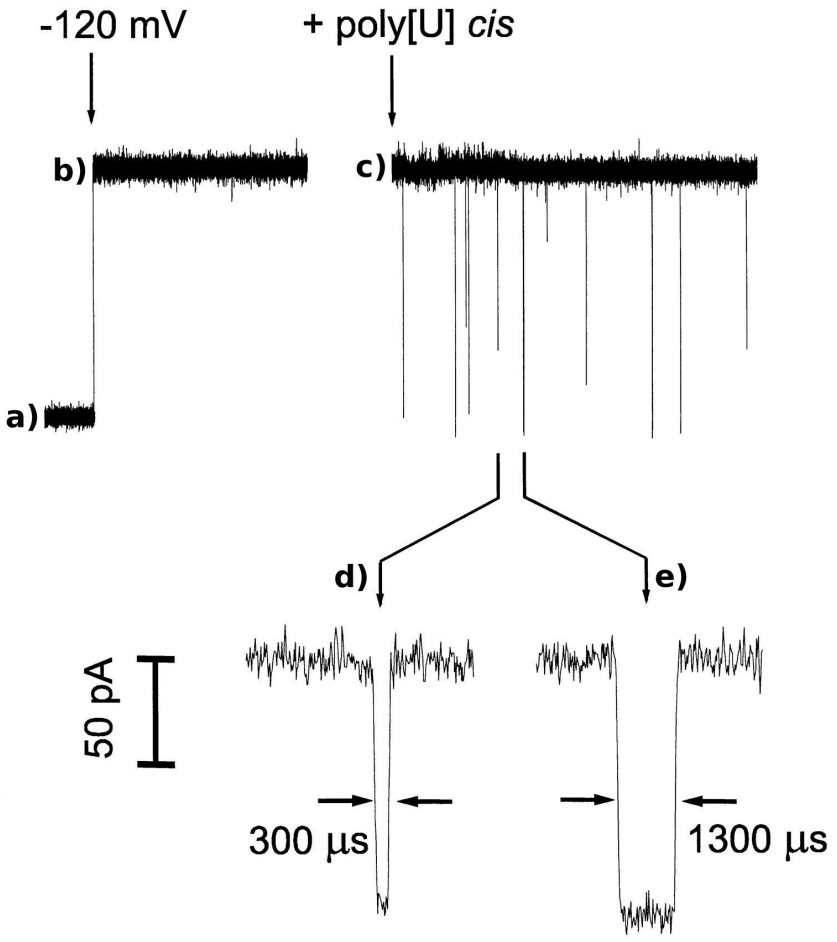


FIGURE 1.4: A single channel current-trace [37] showing a) No current flow, b) The introduction of current through the open channel, c) The introduction of polymers into the *cis*-reservoir, d) and e) Expanded blockage events displaying a measurable degree and duration of blockage.

and conditions, it could be possible to produce a high resolution current trace capable of distinguishing individual nucleotide bases in a single polynucleotide strand.

1.5 Biological and Synthetic Nanopores

Since the original SCCR experiments by Kasianowicz *et al.* in 1996 [37], research in nanopore current recordings has often been conducted with a biological membrane protein pore α-hemolysin (αHL). This protein is shown in Figure 1.5. αHL is a heptameric, transmembrane protein found in *Staphylococcus aureus* [38]. The αHL pore remains open at all times and has internal dimensions allowing for the passage of single stranded

nucleic acid chains but not double stranded. As reported by Song *et al.* [38], the pore of the assembled αHL varies in diameter from 14 Å to 46 Å as described in Figure 1.5. Half-way into the pore lies a constriction: the narrowest part of the αHL-pore, which is the main bottleneck for the passage of ions and nucleic acid strands.

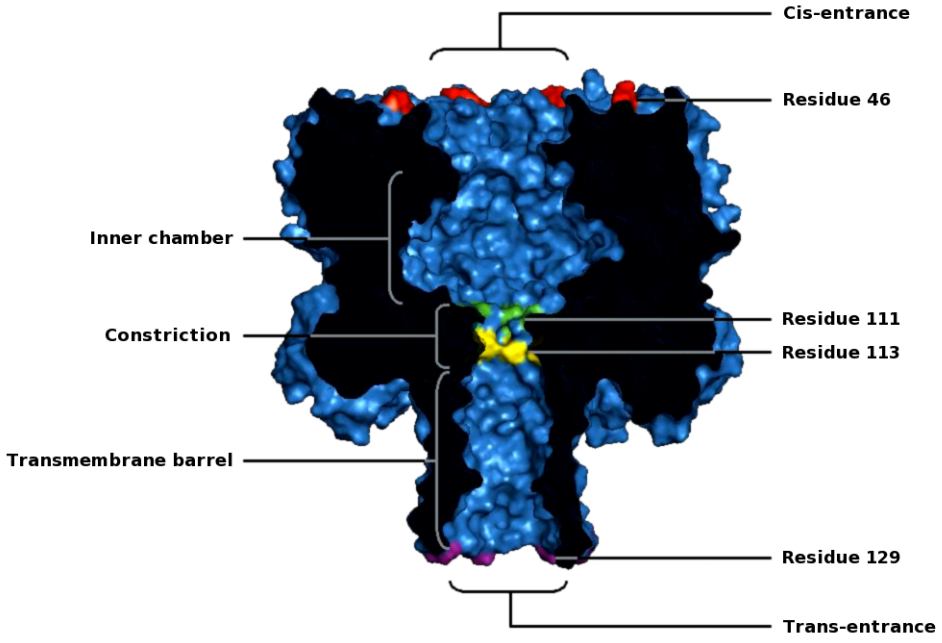


FIGURE 1.5: A cross-section of the heptameric transmembrane form of α-hemolysin. The diameters of the pore features are as follows: *cis*-entrance, 28 Å; inner chamber, 46 Å; constriction, 14 Å; transmembrane barrel, about 20 Å; *trans*-entrance, 24 Å. The height of the entire pore is about 100 Å, while the constriction to the *trans*-entrance measures about 52 Å. Residues of particular interest are also highlighted. Image prepared using PyMOL [39].

Proteins other than αHL have been used in nanopore recording experiments. For example, gramicidin and alamethicin have been used [40, 41]. However, these are ion channels and as such are unsuitable for polynucleotide study due to the internal pore dimensions. Other proteins of suitable dimensions such as OmpG can be ineffective in nanopore recordings due to flexible peptide loops on the pore interior, leading to unstable current flow [42]. MspA from *Mycobacterium smegmatis* is a porin protein that has recently emerged as an attractive option for polynucleotide sensing [43].

Protein nanopores such as αHL possess numerous advantages which make them good targets for experimental and computational research. Firstly, their atomistic structures are usually known from X-ray crystallographic studies. Secondly, their properties may be altered through the relatively simple means of site-directed mutagenesis. It is possible

not only to mutate a residue, but also to alter its properties further by coupling an additional molecule to the pore via, for example, an engineered cysteine and sulfhydryl-reactive chemical component (See Chapter A for examples of this). Additionally, insight learned through the investigation of biological nanopores can be applied to naturally occurring biological systems such as the horizontal transfer of genes across membranes in bacteria. Furthermore, since the crystal structures for protein nanopores are often known, the atomic positions can be converted into structure files for use in computer simulations. There are also various computational tools available for modifying these structure files to account for mutations and other alterations.

It is noted that synthetic nanopores have also been generated for the purpose of nanopore recordings. While of lesser relevance for biological systems, synthetic nanopores present several advantages of relevance for the biotechnological application of DNA sensing and sequencing. Firstly, they have the potential to be significantly more robust than protein nanopores, which may be destroyed or disrupted by high transmembrane potentials and extremes of pH and temperature. One can envisage such robustness being of tremendous value when applied to large scale sequencing of DNA. Not being limited to biological molecules, synthetic nanopores may be constructed from a wider variety of materials. This can grant them interesting properties such as electrical conductance, which, for example, allows electrical measurements to be made directly at the pore [44].

There has been considerable progress in engineering and harnessing synthetic nanopores over the past decade. Using Si_3N_4 as an example material, Li *et al.* demonstrated the sculpting of synthetic nanopores from thin insulating solid-state membranes with highly focused low-energy ion beams [45]. Other groups have sculpted nanopores using electron beams [46, 47]. Iqbal *et al.* manufactured a solid-state nanopore to contain a probe of hair-pin loop DNA which is able to selectively transport short segments of single stranded DNA across the pore [48]. Using an SiN pore, Heng *et al.* discriminated between single-stranded and double-stranded DNA, the latter being unable to translocate through the pore [47]. Most recently, Merchant *et al.* have developed 1-5 nm thick graphene membranes with electron-beam sculpted nanopores from as low as 5 nm in diameter. Since graphene is an electrical conductor, electronic sensing can be performed directly at the pore [44]. Nevertheless, solid state pores are challenging to computationally simulate as the forces of the material must be properly paramaterised. Since the molecular forces

of the twenty naturally occurring amino acids have already been paramaterised, this step is likely not needed when simulating protein nanopores.

Protein nanopores are to be the focus of the investigations in this thesis, rather than synthetic nanopores. This is because of the biological relevance of model protein systems, the more facile computer simulations due to the often known crystal structures, and the easily altered properties through site-directed mutagenesis. Extensive examples of protein nanopore current recordings are presented in the subsequent section. These protein nanopore current recordings provide us with key experimental observations that we can aim to recreate using simulations, giving a means of investigating the validity of the simulated system.

1.6 Protein Nanopore Current Recordings

Over the past 14 years, considerable progress has been made in retrieving information from nanopore current readouts. Even at its earliest stage, results from Kasianowicz *et al.* showed distinction between chains of different length and indicated the possible discrimination of translocation orientation ($5'$ or $3'$-end leading the translocation) [37]. In 1999, Akeson *et al.* used this technique to show clear distinction between homonucleic acid chains by comparing the extent of current blockage and the translocation time observed [49], see Figure 1.6. Signature current-traces of copolymers were also observed by Akeson *et al.*, opening up the possibility of inter-polymer base distinction [49]. Meller *et al.* performed αHL nanopore recordings to show distinction between homopolymers of similar length and composition but differing in sequence. They found that the duration of a translocation event scales with temperature (T) as T^{-2}, laying the foundations for improved SCCR conditions [50].

Several research groups have analysed and adapted αHL nanopore recordings to give further information on the translocation process. Butler *et al.* performed work to show that the signature shape of copolymer traces could be used to indicate the orientation of threading [51]. Meller *et al.* examined the nanopore recordings of DNA strands of various length to determine the velocity of translocation events, concluding that the translocation speed is constant when the DNA is above a certain length [52]. Mathé *et al.* used SCCR to study the kinetics of the unzipping of DNA hairpins [53].

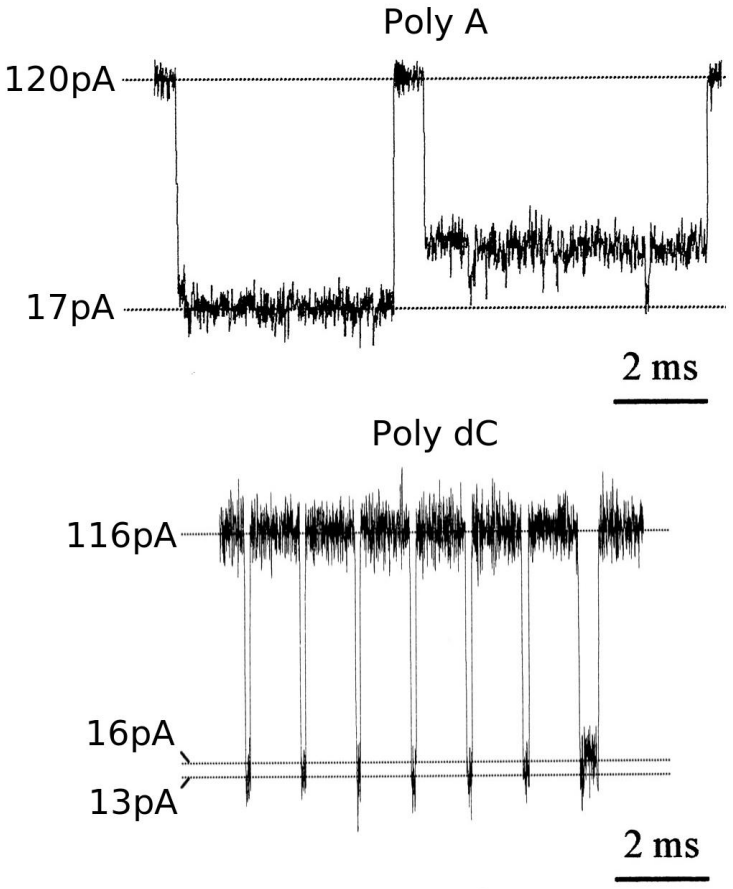


FIGURE 1.6: SCCR trace by Akeson *et al.*, comparing translocation times of poly(dC) and poly(A) chains [49]. The translocation time for poly(dC) is many times shorter than that of poly(A). Note that the length of the poly(A) chains were 175 ± 50 bases, and the poly(dC) chains were 100 bases in length.

Techniques using modified αHL pores have so far yielded promising results. By coupling a single stranded DNA oligomer within the lumen of an αHL mutant, a sequence sensitive SCCR response was developed in 2001 by Howorka *et al.*, whereby a single stranded DNA chain experiences an increased translocation time if the sequence is complementary to the oligomer coupled to the protein [54]. In 2005, Astier *et al.* harnessed an αHL pore equipped with a positively charged molecular adaptor (heptakis-(6-deoxy-6-amino)-β-cyclodextrin) to increase steric and charge effects on the translocating molecules. Using this pore they were able to show clear distinction in the current-blockage events of translocating ribonucleoside and deoxyribonucleoside 5'-monophosphates, differentiating adenine, cytosine, guanine and thymine-based monophosphates [55]. Currently at the forefront of genetic sequencing using nanopores is work performed by Clarke *et al.* [34]. By covalently attaching a cyclodextrin adaptor molecule to the inside of an αHL protein

nanopore, they have demonstrated the ability to identify single nucleotide molecules with an average accuracy of 99.8%. The research group describes the potential in the future to harness an exonuclease enzyme within the experimental setup in order to sequence polynucleotides.

Higher sensitivity information has been retrieved from SCCR-experiments using modified nucleic acid chains. Using DNA hairpins, the discrimination of 5′ and 3′-threaded nucleic acid translocation was substantiated by Mathé *et al.* [56]. Using the observation that 5′-led translocation was different to that of 3′-led, Wang *et al.* produced data showing that phosphorylation of the 5′-end of polynucleotide increased the chance of the 5′-end leading its translocation [57]. This is likely due to the phosphorylation adding negative charges to the 5′ terminus, resulting in an increased chance of that end being captured by the electrical bias across the pore. Ashkenasy *et al.* also used DNA terminated with hairpins, in this case ssDNA was threaded into αHL then the translocation halted due to the sterics of the hairpin. The current flow was at this time strongly dependant on the nucleotide found 20 bases from the cis-entrance into the pore because it was located at the pore constriction. By changing a single nucleotide at this point in the DNA chain, a direct effect on the current trace was observed, leading to single nucleotide discrimination in static DNA strands [58].

1.7 Aims

As demonstrated in this chapter, it would be of considerable value to investigate the mechanisms of the translocation of polynucleotides through protein nanopores on a microscopic level. By producing a solid groundwork of knowledge and understanding, one could better engineer modifications to synthetic and protein nanopores in order to improve the resolution and sensitivity of SCCR output.

The choice of protein and polynucelotides for the investigations in this thesis are guided by the desire to test free energy and translocation algorithms. Firstly, by focusing on the αHL protein-nanopore translocation system, a large set of relevant experimental data is available to draw enlightening comparisons from. Secondly, by examining poly(A) and poly(dC) translocation, the experimental comparisons of the two polynucleotides

presents us with a target observation for us to aim to reproduce, particularly comparing how well our chosen algorithms achieve this. In addition, the findings from such investigations can be applied to a variety of translocation systems involving nanopores, pertaining relevance to biological systems.

Chapter 2

Molecular Dynamics Simulations and Free Energy Calculation Methods

2.1 Computer Simulations

Computer simulations of molecular models can give insight into microscopic processes such as nanopore translocation to further our understanding of experimental observations, potentially providing the basis for new or refined experimental approaches. Recent advances in genetic sequencing using nanopores [34] implicate a fast and cheap method of sequencing, given suitable refinement. A greater understanding of the microscopic translocation factors could greatly aid in thwis regard, which may be achieved through simulation. Given modern resources one cannot arbitrarily replicate nature in simulations to a sub atomic level of precision, however there are techniques for replicating nature in probabilistic detail, atomic detail, even sub-atomic quantum detail [59].

Molecular dynamics (MD) simulations calculate the Newtonian equations of motion of all particles in the system based on the forces present and recalculate the positions of the particles for a small increment in time, usually 1 or 2 femtoseconds. The deterministic time-evolution of the system is calculated in this way, the accuracy of which depends on the level of detail in the model and interactions supplied and allowed. Due to current limits in the timestep of each recalculation, there are limits on the timescales that can be

simulated in atomistically detailed MD. Such limits are typically in terms of nanoseconds, but with the increasing availability and scope of high performance computing facilities, there are instances of microsecond timescales [60, 61] and even millisecond timescales [62, 63]. Since MD simulations reproduce the deterministic time-evolution, they allow for the calculation of the dynamics of the system.

Monte Carlo molecular simulations deal with the statistical mechanics of a system, generating configurations based on probability and measuring the average of desired quantities. This method allows processes which represent a long physical time to be replicated in simulations as it does not follow time evolution in a deterministic way. Trajectory information obtained from Monte Carlo simulations is approximate due to the dependance on statistical mechanics.

Simulations can also be performed at the quantum sub-atomic level. In MD simulations, the model of interactions between the atoms of the system is supplied as the simulation's input, which requires prior knowledge about the interactions of the system. *Ab initio* MD on the other hand use the laws of quantum mechanics of electron and neutron physics to calculate atomic interactions. The main value in *Ab initio* MD is that bond breaking and bond forming is accounted for, unlike in classical MD.

For nanopore translocation we are primarily concerned with the dynamics of the system in atomic detail, which, given unlimited power and resources, would be ideally represented by *ab initio* MD, but given the size of the system to be studied and available resources, classical MD simulations will provide a computationally feasible approach in sufficient detail.

2.2 Molecular Dynamics Simulations with NAMD

For the problem of nucleotide translocation through protein nanopores we have chosen to use MD simulations, allowing atomistic detail, dynamic information, and reasonable timescales to be explored.

The computational complexity of MD algorithms have a lower bound order of N^2 where N is the number of atoms in the model. Given that a nanopore translocation model will typically comprise of more than 250,000 atoms, simulations of such models are highly

computationally expensive. Single processors are very limited in power, to perform a simulation of the relevant magnitude it is necessary to harness many processors in parallel. So for example, a simulation that takes 12 hours to complete on 256 processors would hence take 3072 hours (256 x 12) on a single processor. Therefore the MD code used must be designed with parallel processing capabilities.

NAMD (NAnoscale Molecular Dynamics [64]) is a MD code suitable for simulating nanopore translocation. The code is designed for high-performance large-scale simulations on parallel processors, which makes it ideal to simulate the time-evolution of a large atomistically detailed model such as that of nanopore translocation. There are other parallelised MD programs such as GROMACS [65] or Desmond [66]. NAMD, aside from being available and in a stable form at the start of this project, offers more features than comparable MD codes (such as performing adaptive biasing force simulations) and many more systems have been simulated using the NAMD code, from which we can draw direct comparisons to our work. Currently, for the purposes of the projects covered in this thesis, alternative MD codes offer no significant advantages over NAMD.

The interactions of the system that MD simulations require as input are described by a force-field, such as CHARMM [67] or AMBER [68]. The parameters of such force-fields are determined through high-level *Ab initio* quantum mechanical calculations, and often refined when experimental observations are failed to be reproduced by the MD models using the force-field in question. An error in the force-field for a particular type of interaction can have significant repercussions in the conformations and behaviour of the MD model. For example, in one instance, an error in the dihedral energy terms of AMBER led to the over-stabilisation of α-helices, highlighting the dihedral terms as parameters in need of refinement [69].

NAMD is able to use both the CHARMM and AMBER force-fields. In an atomistically detailed MD simulation, every atom is described by a model force-field which accounts for the interactions between the atom and the system. NAMD uses contributions to the potential energy function from the following interactions:

$$U_{total} = U_{bond} + U_{angle} + U_{dihedral} + U_{vdW} + U_{Coulomb} \tag{2.1}$$

U_{bond}, U_{angle} and $U_{dihedral}$ describe the stretching, bending, and torsional bonded interaction respectively. U_{vdW} describes the nonbonded van der Waal's forces, which is approximated by a Lennard-Jones 6-12 potential. $U_{Coulomb}$ describes the nonbonded electrostatic interactions.

NAMD allows the use of periodic boundary conditions to avoid surface effects at the boundary of simulated systems. By using periodic boundary conditions, the particles in a cell are replicated to infinity by periodic translations, therefore a particle which leaves the cell on one side is replaced by a copy entering the opposite side. The particles in the system are subject to the potential of all the atoms in the cell as well as the potential of the atoms in the repeating cells. Periodic boundary conditions are particularly beneficial as the effectively infinitely sized system requires only the atoms of one cell to be represented in the MD code.

The influence of potential from repeated cells in the system is only of concern for the van der Waals forces U_{vdW} and electrostatic integrations $U_{Coulomb}$, and are computationally unfeasible to calculate the exact long-range interactions. U_{vdW} is approximated by spatially truncating the interactions at a user-defined cutoff distance. For $U_{Coulomb}$, the particle-mesh Ewald (PME) system is applied. PME interpolates the charges of the system to the nodes of a three dimensional grid, so that when a charge is considered by the system, the electrostatic interactions are calculated with respect to the node points which are given weighting functions according to the distance to the charge. This fast numerical method allows the long-range electrostatic interactions to be calculated.

2.3 Free Energy Calculation Methods

With MD selected as the appropriate class of simulation for the nucleotide-nanopore system, and NAMD as the selected MD code, the mode of analysis must be established. Specifically, it must be established which system properties are the most meaningful, and how such properties may be extracted from simulations.

It is desirable to examine the free energy behaviour of chemical processes in order to fully understand them. Through knowledge of the free energy difference associated with a change of state, it is possible to establish stable states, their thermodynamics properties, the kinetics of transitions between states, and it is possible to infer how stable states are

altered by external conditions. Such changes of state include protein mutation, protein-ligand binding, conformational changes, or molecule translocation. It is both possible and useful to calculate experimental free energy changes, and there has recently been a considerable amount of research dedicated to measuring the free energy of simulated systems. The use of simulated systems is pertinent to free energy calculations due to the high control over atomic level perturbation and analysis. As a result, there are numerous ways to approach the problem.

Thermodynamic free energy is the useful work that may be extracted from a closed system. It is a function of the energy of a system, subtracted by an entropy-temperature term, thus it defines the capacity for temperature induced events to occur within a closed system. In the canonical ensemble (constant temperature and volume), the free energy is termed the Helmholtz free energy, A. In the isothermal-isobaric ensemble (constant temperature, constant pressure), it is termed the Gibbs free energy G.

The Helmholtz free energy is defined as:

$$A = U - TS \tag{2.2}$$

Where U is the internal energy (kinetic plus potential energy), T the temperature, and S the entropy. The Gibbs free energy is similar but with an additional pV term where p is the pressure and V the volume:

$$G = A + pV \tag{2.3}$$

In many cases the Gibbs and Helmholtz free energies are interchangeable with respect to the end result of derivations and often in their numerical value (particularly as the system size increases). The Helmholtz free energy will be used for the remainder of this chapter.

The Helmholtz free energy can also be expressed in terms of the partition function, Q:

$$A = -\frac{1}{\beta} \ln Q \tag{2.4}$$

Where β is the inverse temperature, $1/k_B$. Here, the fundamental connection between thermodynamics and statistical mechanics is illustrated, where the calculation of the free energy is equivalent to estimating the partition function. Since we are dealing with the examination of changes in state, we are more focused on free energy differences such that:

$$\Delta A = -\frac{1}{\beta} \ln \frac{Q_1}{Q_0} \tag{2.5}$$

Where Q_0 and Q_1 are the partition functions corresponding to states 0 and 1 respectively, and ΔA being the free energy difference which arises from the transition between the two states. The vast majority of free energy difference problems can be framed in terms of Equation 2.5. Two major groups of free energy calculation which branch from this point are free energy perturbation and thermodynamic integration, both are commonly used in simulations.

2.3.1 Free Energy Perturbation

Free energy perturbation (FEP) theory [70] follows from Equation 2.5. An initial state, 0, is used as the reference problem, then the target state is represented in terms of a perturbation to the reference system. The Hamiltonian of the target state is thus calculated as the Hamiltonian of the reference state plus the perturbation term:

$$\mathcal{H}_1(x, p_x) = \mathcal{H}_0(x, p_x) + \Delta\mathcal{H}(x, p_x) \tag{2.6}$$

Where $\mathcal{H}$ is the hamiltonian of an N-particle system as a function of $3N$ Cartesian coordinates, x, and their momenta, p_x. When the partition functions of Equation 2.5 are substituted with Hamiltonian functions, the formula cancels out to:

$$\Delta A = -\frac{1}{\beta} \ln \langle exp\left[-\beta \Delta\mathcal{H}\left(x, p_x\right)\right] \rangle_0 \tag{2.7}$$

This part of the derivation is dealt with in more detail in Reference [71]. Here, $\langle \dots \rangle_0$ indicates an ensemble average over configurations sampled from the reference state. This is the key idea behind FEP, ΔA can be estimated from sampling only equilibrium

configurations of the reference state. The energy of the target state in relation to the reference state is then computed from those reference state equilibrium configurations.

2.3.2 Thermodynamic Integration

Thermodynamic Integration (TI) [72, 73] is a method which involves the integration of the derivatives of the free energy with respect to an order parameter, λ, along a transformation path. Here, $d\Delta A/d\lambda$ is calculated, rather than ΔA. Thus Equation 2.4 is differentiated to give:

$$\frac{d\Delta A}{d\lambda} = -\frac{1}{Q(\lambda)}\frac{\partial Q}{\partial \lambda} \tag{2.8}$$

Similarly to the case with FEP, if the partition functions are substituted for Hamiltonians with λ as a parameter, terms cancel out (see Reference [71]) to give:

$$\frac{dA}{d\lambda} = \left\langle \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda} \tag{2.9}$$

From this equation we can see that ΔA between state 0 and state 1 is calculated through the integration of the average derivative of the Hamiltonian with respect to λ in the range λ_0 and λ_1

2.3.3 Sampling and Potentials of Mean Force

The phase space of a system is a multi-dimensional space in which all possible configurations of that system are represented. Each particle in a system has degrees of freedom associated with it: defining the positions and momenta of that particle. Each axis in the phase space represents a degree of freedom. The configurations represented in the phase space have energy terms and probability densities (at $t = \infty$) associated with them. A free energy landscape is a multi-dimensional space where at least one axis represents the free energy and the other axis or axes each represent a reaction coordinate (a one dimensional function representing a reaction pathway). Each point on the free energy landscape is a function of the phase-space configurations represented by that point, and the average of their associated free energies: which is weighted according

to the probability density (at $t = \infty$) of those phase-space configurations. The goal of many simulations studies is to determine profiles of such energy landscapes.

The potential of mean force (PMF) is defined as the free energy profile along a well defined reaction coordinate, ξ. The reaction coordinate may be, for example, a translocation distance or a chemical reaction path. In practice, there are key considerations to be made when sampling the phase space of the reaction coordinate in order to produce the PMF. For ergodic systems, the probability of visiting a point in phase space converges to a limiting value as time approaches infinity. At $t = \infty$ each point in phase space will have been visited proportionally to its probability density function, and so a free energy estimate based on this data will be precise. Simulations will of course be finite in length and so the free energy calculated from sampling simulation data will only be an estimate. However, for finite sampling, some key configurations may not be visited at all. This could be due to, for instance, a dependance on the initial conditions, states separated by high barriers in the free energy landscape, or entropy bottlenecks. Such systems are described as quasi-nonergodic due to their non-ergodic appearance.

There are various methods aimed at improving the sampling of reaction coordinates. Stratification and umbrella sampling are two such methods and are illustrated in Figure 2.1. Using stratification the reaction coordinate is split into intervals, these intervals are much narrower than the entire reaction coordinate, reducing trapping from energy barriers or entropy bottlenecks. Additionally, as in the case of umbrella sampling (US) [74, 75], biasing potentials may be applied in order to encourage exploration on or past high energy barriers. The biased potential is removed from the subsequent free energy calculation. Choosing such intervals and biasing potentials requires an *a priori* knowledge of the free energy landscape.

2.3.4 Adaptive Biasing Force

Adaptive biasing force is an advanced form of US which estimates the free energy landscape during the sampling simulation and applies biasing forces which allows for effective sampling. In US, an external potential is applied to allow for the exploration of higher energy configurations and the states that they separate. Here, the free energy along a chosen reaction coordinate (ξ) is given by:

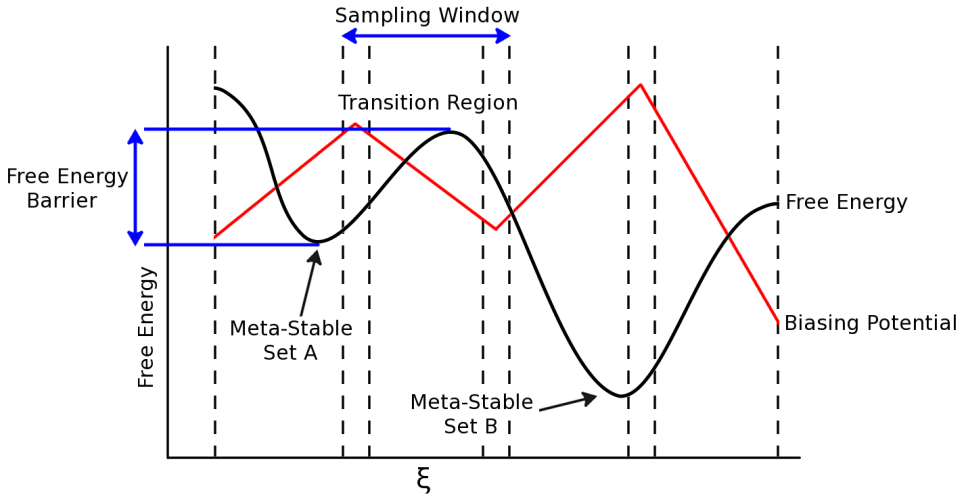


FIGURE 2.1: Illustration of stratification and umbrella sampling. The figure shows free energy plotted against the reaction coordinate, ξ. The reaction coordinate is split into overlapping sampling windows, reducing the number of free energy barriers per interval. The biasing potential (red), designed to overcome the free energy barriers, is applied in the form of a linear ramp within each sampling window.

$$F(\xi) = \frac{1}{\beta} ln \mathcal{P}_{\xi} - \mathcal{U}_{bias} + F_0 \tag{2.10}$$

Where $F(\xi)$ is the free energy of the state at a particular value of ξ. $\mathcal{P}_{\xi}$ is the probability density of finding the system at ξ. $\mathcal{U}_{bias}$ is the applied external potential. β is the inverse temperature, and F_0 is a constant. If the U_{bias} is tuned to be the exact opposite of the free energy, $-F(\xi)$, then the free energy landscape will effectively be flattened out. This then permits uniform sampling along the reaction coordinate. Knowing the landscape of $-F(\xi)$ implies an *a priori* knowledge of the free energy landscape. If the U_{bias} deviates from $-F(\xi)$ then the reaction coordinate will be poorly sampled. Thus when using US for quantitatively new problems, good sampling of the reaction coordinate is very difficult.

Adaptive biasing force (ABF) is a method of reaction coordinate sampling which applies a continuous biasing force, which is tuned during the simulation to the cumulative estimate of the free energy landscape. Uniform sampling can thus be achieved by the on-the-fly calculation of $F(\xi)$ and the implementation of this information in the form of an external bias. The implementation of ABF, therefore, requires no knowledge of the free energy landscape prior to the simulation.

The algorithm to achieve the adaptive biasing force was developed by Darve and Pohorille [76, 77] and implemented into the NAMD code by Hénin and Chipot [78].

The derivative of the free energy with respect to the reaction coordinate can be expressed in terms of configurational averages at constant ξ:

$$\frac{dF(\xi)}{d\xi} = \left\langle \frac{\partial \mathcal{V}(x)}{\partial \xi} - \frac{1}{\beta}\frac{\partial ln|J|}{\partial \xi} \right\rangle_{\xi} = -\langle f_{\xi} \rangle_{\xi} \tag{2.11}$$

The $\partial \mathcal{V}(x)/\partial \xi$ term represents the physical forces exerted on the system, as derived from the potential energy function, $\mathcal{V}(x)$. The $(1/\beta)(\partial ln|J|/\partial \xi)$ term represents a geometric correction where $|J|$ is the determinant of the Jacobian for the transformation from generalised to Cartesian coordinates, this term accounts for difference in phase space availability as the reaction coordinate varies. $\langle f_{\xi} \rangle_{\xi}$ is the average force acting along reaction coordinate ξ, derived from instantaneous force components f_{ξ}.

2.3.4.1 The Implementation of the Biasing Force

The biasing force of the ABF method is calculated as the negative of the average force acting along the reaction coordinate, $\langle f_{\xi} \rangle_{\xi}$, derived from instantaneous force components f_{ξ}. For its practical implementation, f_{ξ} is accumulated in small bins along the reaction coordinate, $\delta\xi$, to provide an estimate of the change in free energy with respect to the reaction coordinate, $dF(\xi)/d\xi$. The adaptive biasing force, $\mathbf{f}^{ABF}$ is thus defined as:

$$\mathbf{f}^{ABF} = \frac{dF(\xi)}{d\xi} = -\langle f_{\xi} \rangle_{\xi} \tag{2.12}$$

Here, $F(\xi)$ is the current estimate of the free energy as a function of the reaction coordinate. The calculated biasing force, $\mathbf{f}^{ABF}$, is introduced into the equations of motion during a simulation. With the biasing force applied, the overall forces acting on the biased atom(s) along the reaction coordinate within a bin average to zero over time, since the biasing force approximately opposes the energetic barriers to translocation. As the applied biasing force roughly matches to the free energy landscape, evolution along the reaction coordinate is largely governed by self-diffusion properties.

The instantaneous force, f_ξ, fluctuates to a high degree, thus the average force will initially take inaccurate and physically meaningless values. Therefore, the biasing force is not implemented within a particular bin until a threshold of force measurements has been accumulated, the biasing force is then introduced gradually via a linear ramp. The threshold value, ζ, is an input parameter in units of timesteps and its best chosen value is dependent on the nature of the system. The size of ζ aids in the reduction or removal of unwanted non-equilibrium effects, therefore it will need to be larger for systems with slowly relaxing degrees of freedom. For example, those systems with large, flexible translocating molecules. This threshold parameter heavily influences the translocation/simulation time, though it is not the only factor in this regard due to the dependence of the biased atom(s) on self-diffusion.

To keep the atom(s) within the bounds of the desired reaction coordinate, harmonic boundary forces are applied at either end. The position of the boundaries and the value of their force constants are selected as simulation input parameters. The boundary forces are rigidly implemented and may influence the quality of the data near to them [79], therefore it is beneficial to ensure that the distance between the boundaries is as long as possible. However, as we shall see later in Chapter 4, there are benefits to segmenting the full reaction coordinate into several shorter intervals.

The ABF reaction coordinate is often defined in terms of relative atomic positions, that is, the distance between selected atoms or groups of atoms. In the latest implementation of ABF (in NAMD2.71b), the reaction coordinate may be defined in various ways, including bond angles, radii, and RMSD values [80]. For the studies in Chapter 4, the reaction coordinate is most appropriately designated in terms of inter-atomic distances in the z-axis direction. Since the reaction coordinate is relative to two atoms or groups of atoms, the Cartesian coordinates of the moving and reference atoms or groups must be converted into generalised coordinates, as outlined in Section 2.3.4.

The mode of translocation in ABF is described in Figure 2.2. The means of progression of the translocating molecule along the reaction coordinate is important in understanding the observations from the free energy profiles in Chapter 4. The Figure shows a time sequence plotted against translocation distance represented by illustrations (a)-(f); below them is a free energy profile that determines these events, also plotted against

translocation distance. (a) The biased atom (blue circle) is within the first bin of the reaction coordinate, the instantaneous force components acting on the biased atom in the z-direction are being accrued in order to calculate the biasing force. The bias is not being applied to the atom at this time, as the threshold parameter, ζ, has not been reached. The bin is coloured yellow to indicate the force calculations being done. The atom may not move outside of the boundaries of the reaction coordinate due to a harmonic bias applied at the ends. (b) With ζ surpassed, the calculated biasing force roughly opposes the free energy barrier to moving into the next bin, allowing translocation by diffusion. From the free energy profile we can see that the barrier to translocation is small, and so only a small biasing force is required to aid translocation, indicated by the small red arrow. The atom has diffused into the proceeding bin and the instantaneous force samples are now being accrued there. The red arrow in the green box indicates that the previous bin has been sufficiently sampled, thus a biasing force will be applied if the atom moves back into that bin. (c) Due to the shape of the free energy profile, a larger biasing force is calculated for the second bin (large red arrow). (d) Because the biasing force attempts to oppose the free energy landscape, the atom may move back into previous bins as indicated here (this may also occur before ζ has been reached). The instantaneous force components are accrued even after ζ has been reached (with a subtraction for the currently applied biasing force), and the biasing force is recalculated as indicated by the yellow arrow. (e) The atom has diffused into the proceeding bin where the biasing force was calculated in (c), due to the free energy landscape the biasing force tends backwards, but as is opposes, not exceeds, the free energy landscape, it may continue to diffuse forwards. (f) The biased atom has reached the end of the reaction coordinate, it may diffuse backwards across the reaction coordinate if the simulation continues.

2.3.4.2 Examples of ABF Applications

ABF has been applied to various systems in order to obtain the free energy profile of an event, some of the key examples are highlighted in this section.

In 2005, Chipot and Henin performed the first application of ABF to a biomolecular system [79]. The authors produced profiles of the free-energy landscape for the stretching of deca-alanine in classical MD simulations, providing a thorough exploration into the

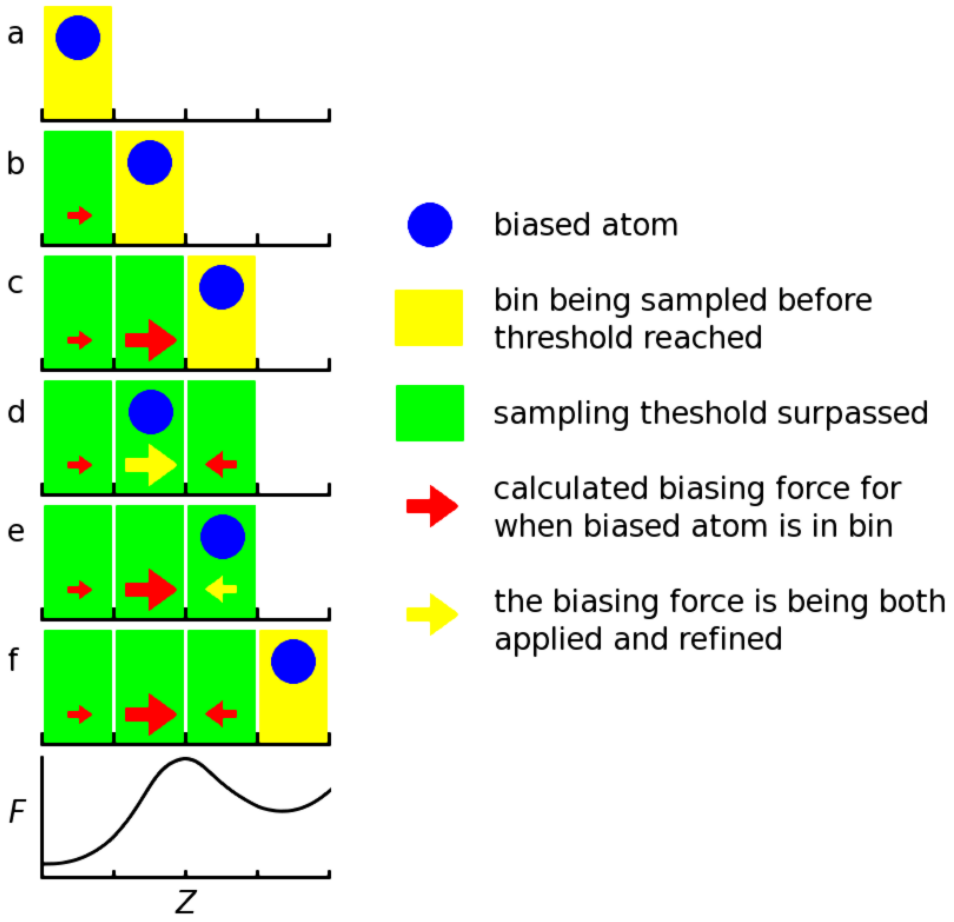


FIGURE 2.2: Illustration of the ABF methodology. The x-axis represents the translocation reaction coordinate, which is the z-axis distance down the pore, the reaction coordinate is discretised into bins. The ABF methodology can be summarised through the time sequence represented by illustrations (a)-(f) and the free energy profile at the bottom. (a) The biased atom (blue circle) is within the first bin of the reaction coordinate. The biasing force is being estimated (yellow rectangle). (b) The biased atom has been sampled more than the threshold, ζ, in the first bin and the biasing force (red arrow) has allowed to atom to move into the subsequent bin. (c) The biased atom has been sampled more than the threshold ζ in the second bin and has moved to the subsequent bin. The biasing force to move the atom from the second bin was larger than in the first, due to the free energy landscape. (d) The biasing force in the third bin has allowed the biased atom to diffuse in either direction. In this case it has returned to the second bin. The biasing force is being continually refined (yellow arrow). (e) The biased atom has moved back into the third bin. The biasing force is being continually refined. (f) The biased atom as reached the end of the reaction coordinate, it may continue to diffused backwards across the reaction coordinate if the simulation continues.

effects of changing ABF-specific parameters. The resulting data was compared to that of an unbiased (but otherwise identical) 50 ns reference simulation in order to indicate deviation from reliable values, and the free energy profiles were constructed after averaging the data from ten samples. The system was simulated *in vacuo*, so the properties of the system are intrinsically different to the large, solvated and ionised systems explored in this thesis. While the deca-alanine system is intrinsically different to the polynucleotide-nanopore system, it indicates key considerations that must be made in setting up ABF simulations for other models. The NAMD2.6 build used at the time of the paper's writing has since been found to exert the boundary force constants incorrectly; this has been corrected in the ABF implementation of NAMD2.7b1. Nonetheless, the authors observed direct effects in the sample-to-sample variation, the sampling uniformity, and the deviation from the reference simulation when parameters such as the bin size, the boundary force constants and the force measurement threshold were changed [79].

Henin *et al.* used the adaptive biasing force method to translocate glycerol through *Escherichia coli* aquaglyceroporin, GlpF [81]. The same system was explored in 2002 using cv-SMD in conjunction with Jarzynski's equality to produce the free energy landscape of the translocation event [82]. Using cv-SMD, Jensen *et al.* were able to gain insight into the translocation process, revealing the stereoselectivity of the channel and channel-glycerol hydrogen bonding interactions. Using ABF, Henin *et al.* were able to expand on these findings, highlighting the critical role played by intramolecular relaxation on the diffusion properties of the translocating molecule.

Other examples of ABF being used in membrane channel proteins include the extraction of free energy profiles from potassium ion translocation through the Kv1.2 protein ion channel by Treptow and Tarek, gaining insight into the gating mechanism [83]. The authors found agreement with experiment that the region of Val^{478} constitutes the main gate, but expanded on this by determining from free energy profiles that the conduction through the gate is unfavourable when the constriction is smaller than an estimated threshold of 4.55.0 Å. The ABF method was used in a similar way by Ivanov *et al.*, the study examined the barriers to ion conductance in cationic human $\alpha 7$ nicotinic acetylcholine receptor (nAChR) and the anionic $\alpha 1$ glycine receptor (GlyR) [84]. Lamoureux *et al.* produced the free energy profiles from water molecule translocation in the AmtB ammonium transporter, exploring the accessibility of the water molecules to the hydrophobic pore [85]. The free energy profiles revealed that there were no significant

hydrophobic barrier for chloride translocation, this allowed them to conclude that the overall electrostatics and the presence of charged residues at the entrance and exit of the channels were sufficient to explain the ion selectivity.

The ABF method has also being used in various situations which do not involve membrane channel proteins. Wei and Pohorille used ABF to explain the experimental finding that ribose permeates membranes an order of magnitude faster than arabinose and xylose (its diastereomers) [86]. The simulations were in excellent agreement with experiment, by comparing the molecular models with the transmembrane free energy difference between ribose and and its diastereomers, the difference was attributed to stronger intramolecular interactions between the hydroxyl groups of ribose. In similar cases of using ABF to gain insight into experimentally found trends, Xu *et al.* explored the organisation and assembly of the cellulosome produced by anaerobic bacteria [87]. The study examined the affinity of protein cohesin domains with dockerins of cellulosomal enzymes, using ABF to produce free energy profiles of cohesin-dockerin complexation, revealing a high-free energy barrier against dissociation. Cai *et al.* explored the inclusion of hydrocortisone, progesterone, and testosterone into the cavity of β-cyclodextrin [88]. Rodriguez *et al.* examined the docking of 6-amino-6-deoxy-2O-carboxymethyl-β-cyclodextrins (JCD) in aqueous solutions, which form nanotubes [89]. Vaitheeswaran and Thirumalai used ABF to estimate the free energies of interaction between amino acid side chains in cylindrical hydrophobic synthetic nanotubes, the results having applications to peptide stability [90]. Dehez *et al.* examined the binding of ADP in the membrane protein AAC (mitochondrial ADP/ATP carrier) [91].

Most recently, Henin *et al.* have expanded the NAMD implementation of the ABF method to allow the production of multidimensional free energy landscapes, being previously restricted to one dimension [80]. In this work, they produce 2- and 3-dimensional free energy landscapes from reaction coordinates which include inter-atomic distances, radii, and RMSD values. They study several systems to this end: The conformational equilibrium of N-acetyl-N′-methylalaninamide, chloride ion permeation across a peptide nanotube, and the structure of the opioid growth factor met-enkephalin in aqueous solution. They also return to their previous study [79] of the conformational landscape of deca-alanine, the free energy landscape now being expanded to 2- and 3-dimensions.

The ABF method is explored for use in the nucleotide-nanopore system, in Chapter 4.

2.4 Non-equilibrium and Steered Molecular Dynamics

In calculating free energy differences, there is a general formalism which is followed, namely the transformation between two states. This transformation can be performed infinitely slowly as in TI, or infinitely fast, as in FEP. Methods for performing the transformation at finite speed exist, and must account for the introduction of non-equilibrium, which is inherent to such an approach. Two such methodologies include use of the Jarzynski identity and the Crooks relation.

The use of non-equilibrium simulations is easily justified for the nucleotide translocation through a nanopore problem. MD simulations of many atom systems in full atomistic detail are computationally expensive. For detailed simulations of the SCCR system, microsecond and millisecond timescales would need to be represented. Rather than making approximations in the detail of the model, one approach is to use alternative forms of translocation to a transmembrane current.

Steered Molecular Dynamics (SMD) provides a means of retrieving more data in a smaller timescale while keeping atomistic detail. This comes at the cost of the system being at non-equilibrium. Approximations do not necessarily have to take the form of model simplifications, MD programs such as NAMD [92] allow simulated processes to be steered by introducing non-equilibrium forces. Here for example, an atom can have a directional force applied to it, causing the atoms and anything coupled to the atoms to move with the direction of the force. The benefit of this is that a significant degree of movement can be induced in a relatively short timeframe, and the full atomistic dynamics of the process can be examined.

Applying SMD to threading nucleic acid chains through nanopores, the translocation process can be replicated in a atomistic model within a timeframe that can be simulated. By pulling the nucleic acid strand at constant velocity (known as constant velocity steered MD or cv-SMD [92]), a translocation of known distance can be performed. In constant velocity SMD, an atom or the centre of a group of atoms is harmonically restrained to a point in space that is shifted in the chosen direction. The harmonic restraint can be thought of as a spring attached to a dummy atom, the strength of the restraint is given by a force constant k. The simulation outputs the force in piconewtons (pN) experienced by the spring in the direction of pulling (the reaction coordinate).

An alternative to cv-SMD is constant force SMD (cf-SMD), here the constraint point is moved to keep the force on the restraint at a constant, resulting in a variable velocity. This is advantageous when the applied force must remain limited, but one loses control over translocation time, hence the length of the simulation. Therefore cf-SMD is not applicable for these simulations.

2.4.1 Jarzynski's Equality

Free energy is an equilibrium property, while cv-SMD introduces non-equilibrium forces. Jarzynski's Equality (JE) [93] equates equilibrium free energy to ensemble average of the exponential work of a non-equilibrium process. Allowing a means of extracting the free energy from a non-equilibrium simulation.

Consider a process changing a parameter, λ, of a system from equilibrium point 0 at a time zero, to Λ at time τ. The second law of thermodynamics states the average work done, $\langle W \rangle$, on the system cannot be smaller than the Helmholtz free-energy difference between the initial and final states of λ:

$$\Delta F = F(\lambda_\tau) - F(\lambda_0) \leq \langle W \rangle \tag{2.13}$$

However, Jarzynski determined that the free-energy difference between the two states can be related to the work, W, by the following equality:

$$e^{-\beta \Delta F} = \left\langle e^{-\beta W} \right\rangle \tag{2.14}$$

Here β is the inverse temperature, $\frac{1}{k_B T}$, where k_B is Boltzmann's constant.

2.4.1.1 Derivation of Jarzynski's Equality

Jarzynski's equality was first derived for Hamiltonian systems [93]. Consider a classical mechanics based system (S) in contact with heat bath (B) of constant temperature. The compound (SB) is thermally isolated and evolves according to Hamiltonian dynamics. Under the assumption that the surface energy between the system and the bath is

negligible, the Hamiltonian of the compound can be divided into Hamiltonians of the system and the bath separately:

$$\mathcal{H}_{\lambda}^{SB}(\Gamma,\Theta) = \mathcal{H}_{\lambda}^{S}(\Gamma) + \mathcal{H}^{B}(\Theta) \tag{2.15}$$

Here, Γ denotes the position and momentum phases of the system, while Θ denotes the phases of the bath. The Hamiltonian of the system depends on λ therefore the partition function is factorised as:

$$\begin{aligned} Q_{\lambda}^{SB} &= \int d\Gamma d\Theta exp\left[-\beta\mathcal{H}_{\lambda}^{SB}(\Gamma,\Theta)\right] \\ &= \int d\Gamma exp\left[-\beta\mathcal{H}_{\lambda}^{S}(\Gamma)\right] \int d\Theta exp\left[-\beta\mathcal{H}^{B}(\Theta)\right] \\ &= Q_{\lambda}^{S}Q^{B} \end{aligned} \tag{2.16}$$

The time evolution of the compound SB is determined by the time-dependant Hamiltonian $\mathcal{H}_{\lambda}^{SB}(\Gamma,\Theta)$, where the time dependence arises solely from λ, and the initial and final states can be labelled as (Γ_0,Θ_0) and $(\Gamma_\tau,\Theta_\tau)$ respectively. The compound SB is best described by a microcanonical ensemble as it is thermally isolated. However even when the system is not *macro*scopic, the compound SB is, therefore it is permissible to use a canonical ensemble. Hence the initial state, (Γ_0,Θ_0) may be sampled from the distribution:

$$\frac{1}{Q_0^{SB}} exp\left[-\beta\mathcal{H}_0^{SB}(\Gamma_0,\Theta_0)\right] \tag{2.17}$$

Due to energy conservation, the work done during the process is equal to the difference in the Hamiltonians of the initial and final states:

$$W = \mathcal{H}_{\Lambda}^{SB}(\Gamma_\tau,\Theta_\tau) - \mathcal{H}_0^{SB}(\Gamma_0,\Theta_0) \tag{2.18}$$

Therefore the average of the exponential work, $\left\langle e^{-\beta W}\right\rangle$, is given by the following:

$$\begin{aligned}\left\langle e^{-\beta W}\right\rangle &= \int d\Gamma_0 d\Theta_0 \frac{1}{Q_0^{SB}} exp\left[-\beta \mathcal{H}_0^{SB}\left(\Gamma_0,\Theta_0\right)\right] \\ &\quad exp\left\{-\beta\left[\mathcal{H}_\Lambda^{SB}\left(\Gamma_\tau,\Theta_\tau\right)-\mathcal{H}_0^{SB}\left(\Gamma_0,\Theta_0\right)\right]\right\} \\ &= \int d\Gamma_0 d\Theta_0 \frac{1}{Q_0^{SB}} exp\left[-\beta \mathcal{H}_\Lambda^{SB}\left(\Gamma_\tau,\Theta_\tau\right)\right] \end{aligned} \tag{2.19}$$

The final phase $(\Gamma_\tau, \Theta_\tau)$ can be obtained from the initial phase (Γ_0, Θ_0) by the forward time evolution of the Hamiltonian system, and the initial from the final by the backwards time evolution. Hence, the integration variable (previously (Γ_0, Θ_0)) can be transformed to $(\Gamma_\tau, \Theta_\tau)$. According to Liouville's theorem, the Jacobian of the transformation is unity, giving:

$$\begin{aligned}\left\langle e^{-\beta W}\right\rangle &= \int d\Gamma_\tau d\Theta_\tau \frac{1}{Q_0^{SB}} exp\left[-\beta \mathcal{H}_\Lambda^{SB}\left(\Gamma_\tau,\Theta_\tau\right)\right] \\ &= \frac{Q_\Lambda^{SB}}{Q_0^{SB}} = \frac{Q_\Lambda^{S}}{Q_0^{S}} \end{aligned} \tag{2.20}$$

Finally, using Equation 2.16 we get Jarzynski's equality between the work and the Helmholtz free energy difference of the system:

$$\left\langle e^{-\beta W}\right\rangle = \frac{Q_\Lambda^S}{Q_0^S} = e^{-\beta \Delta F} \tag{2.21}$$

2.4.1.2 Calculating the PMF

The raw output of a cv-SMD simulation gives the force (f) in piconewtons experienced by the SMD spring with respect to time and/or extension, from this the energy profile in the form of the PMF must be calculated. In pulling simulations, work is defined as a force-versus-extension integral:

$$W = \sum_{i=1}^{N_S} f_i \Delta x_i \tag{2.22}$$

Where x_i is the distance pulled and N_S is the number of intervals in the sum. Therefore the data can be converted into work with respect to extension. From the work, a rearrangement of Jarzynski's equality can be used to obtain the Helmholtz free energy difference:

$$F_{\lambda(\tau)} - F_{\lambda(0)} = -\frac{1}{\beta} log \left\langle exp \left[-\beta W \left(\tau \right) \right] \right\rangle \tag{2.23}$$

The exponential average work $\langle exp \left[-\beta W \left(\tau \right) \right] \rangle$ is taken over an ensemble of trajectories of the process, which is approximated in MD by performing repeat simulations from initial points at equilibrium. A difficulty in using Jarzynski's equality is that the exponential average work does not converge well due to the system being dominated by small values of work. These configurations are not explored well due to their rare occurrence. Therefore instead of estimating the exponential average directly, the cumulant expansion is employed, which, in the case of a logarithm of an exponential average (as in Equation 2.23) is expanded in terms of the following cumulants:

$$log \left\langle e^x \right\rangle = \left\langle x \right\rangle + \frac{1}{2!} \left(\left\langle x^2 \right\rangle - \left\langle x \right\rangle^2 \right) + \frac{1}{3!} \left(\left\langle x^3 \right\rangle - 3 \left\langle x^2 \right\rangle \left\langle x \right\rangle + 2 \left\langle x \right\rangle^3 \right) + ... \tag{2.24}$$

Hence expanding Equation 2.23 gives:

$$\begin{aligned} F_{\lambda(\tau)} - F_{\lambda(0)} \quad = \quad & \left\langle W \left(\tau \right) \right\rangle + \frac{\beta}{2!} \left(\left\langle W \left(\tau \right)^2 \right\rangle - \left\langle W \left(\tau \right) \right\rangle^2 \right) \\ & + \frac{\beta}{3!} \left(\left\langle W \left(\tau \right)^3 \right\rangle - 3 \left\langle W \left(\tau \right)^2 \right\rangle \left\langle W \left(\tau \right) \right\rangle + 2 \left\langle W \left(\tau \right) \right\rangle^3 \right) + ... \end{aligned} \tag{2.25}$$

Here the series can be terminated at a certain order to give an approximate formula for the Helmholtz free energy difference of the process. While using the cumulant expansion will reduce the sampling error of estimating the exponential average directly, a truncation error is introduced by ignoring higher order terms. In the example of stretching deca-alanine [94], by comparing the irreversible pulling PMF with the exact PMF of the reversible process, the second order cumulant expansion and the exponential average were found to give the best estimates for the free energy, the second order cumulant expansion giving a slightly better estimate.

As shown by Park *et al.* [95], the PMF Φ is related to the free energy F_λ by the following:

$$\Phi(\lambda) = F_\lambda + \frac{1}{2k}\left(\frac{\partial F_\lambda}{\partial \lambda}\right)^2 - \frac{1}{2\beta k}\frac{\partial^2 F_\lambda}{\partial \lambda^2} + O\left(\frac{1}{k^2}\right) \tag{2.26}$$

Where k is the spring constant for the guiding potential. In applying this to SMD simulations, a sufficiently large value of k justifies the approximate equality of $\Phi(\lambda) = F_\lambda$. Therefore using Equation 2.25 on the work done from a series of repeat simulations, the PMF is obtained.

The form of the simulation output data and examples of the data during each state of calculation of the free energy profiles is shown in Figure 2.3. Figure 2.3a shows the force on the harmonic spring connecting the SMD atom to the constraint position moving at constant velocity, plotted against the displacement of the SMD atom. Figure 2.3b shows a local work profile and a local free energy profile. Figure 2.3c shows a global work profile and a global free energy profile.

2.4.1.3 Examples of Constant Velocity-Steered Molecular Dynamics and Jarzynski's Equality

Much work has gone into exploring the validity of Jarzynski's equality when applied to simulated and real experiments. The already mentioned helix-coil transition of deca-alanine by Park *et al.* gave an excellent comparison of PMF reconstructions in irreversible and reversible systems [95]. PMF reconstructions were examined on glycerol conduction through Aquaglycerolporin GlpF by Jensen *et al.*, illustrating the application of JE to large systems and in particular, transmembrane channels [82]. The adiabatic representation (a system not coupled to a bath) was used by Mukamel *et al.* to show that the JE also holds for quantum systems [96]. JE was applied to simulated and real AFM experiments by Hummer *et al.*, which approximate pulling experiments [97]. Douarche *et al.* have applied PMF reconstructions to real experiments on the thermal fluctuations of a macroscopic mechanical oscillator in contact with a heat reservoir [98].

JE has been used to extract free energy profiles from a cv-SMD stretching of deca-Alanine in a vacuum [94]. The free energy profile for irreversible pulling was calculated using the first, second and third order cumulant expansion of JE (discussed in Section 2.4.1.1)

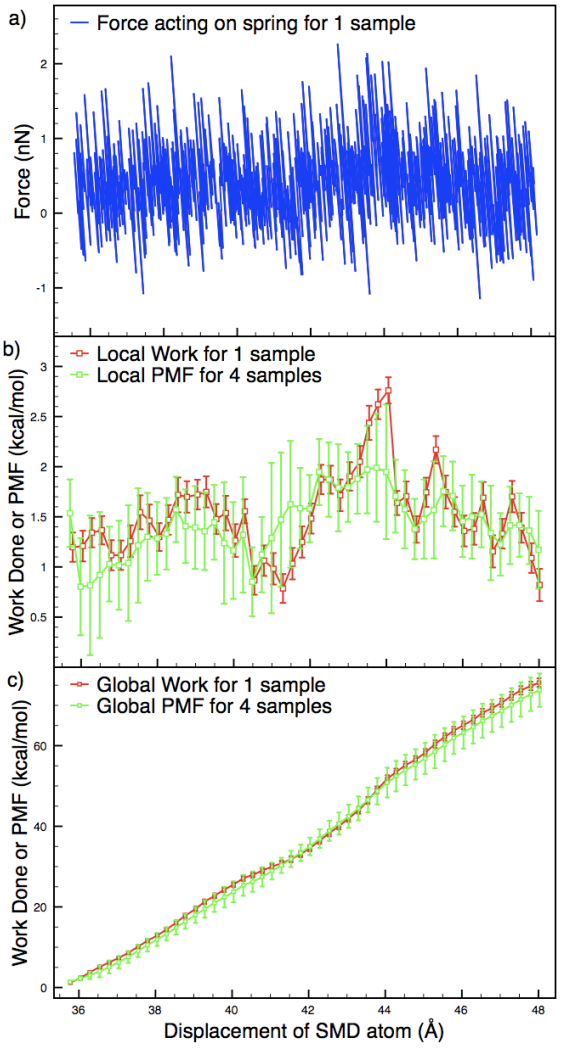


FIGURE 2.3: Example plots at each stage of data processing towards producing a free energy profile. (a) The force on the harmonic spring connecting the SMD atom to the constraint position moving at constant velocity, plotted against the reaction coordinate (z-axis displacement of the SMD atom in Angstroms) for a single sample of dC_{25} translocating through the αHL constriction. (b) The local work profile for one sample (red) and the local free energy profile (green). The local work profile is calculated from the force acting on the spring by converting it into work bins, $f.\Delta x$, where Δx is the width of the bin (0.25 Å). For local profiles each bin represents the work done solely in that bin. The local free energy profile is calculated to a first approximation from the ensemble average of the work done. In this case, the average of four local work profiles is taken. (c) The global work profile (red) and the global free energy profile (green). The work is calculated from the force acting on the spring with respect to extension according to $f.\Delta x$. For global profiles each bin represents the sum of all work bins from the start of the reaction coordinate to that point. The global free energy profile is calculated as the ensemble average of the work done. In this case, the average of four global work profiles is taken. For work profiles the error bars include the binning error, for free energy profiles the error bars include both the binning error and the sample-to-sample error.

as well as the exponential average where the cumulant expansion is not applied. The results were compared to the free energy profile of a reversible pulling of deca-Alanine with consideration made to the standard deviation over the bins and it was found that the second order cumulant gave the best estimate while the first order cumulant expansions gave the lowest standard deviation. It was found that increasing the force constant, k, on the cv-SMD spring increases the statistical error, while it is necessary to approximate a stiff spring is order to control the systematic errors, indicating that care should be taken when selecting this parameter. It was also shown that increasing the speed of the pulling results in irreversible work moving the free energy profile away from the reversible profile. The irreversible work is discounted by JE and was shown to do so at 10 Å/ns pulling, however at 100 Å/ns the irreversible free energy profile did not align within error bars to the reversible profile. The increasing pulling speed was also shown to increase statistical fluctuations, indicating that pulling speed should be minimised where possible.

JE has also been used to reconstruct the free energy profile from a cv-SMD translocation of glycerol through transmembrane protein aquaglycerolporin GlpF [82]. The free energy profile was used to calculate an estimate of the transport time - a kinetic quantity, using an approximate intra-channel friction coefficient. The simulations revealed binding sites and energy barriers in the pore which, for example, led to insight into the mechanism of selectivity filtering that GlpF exhibits.

The use of Jarzynski's equality for the nucleotide translocation through a nanopore problem is presented in Chapter 3.

2.4.2 Crooks' Fluctuation Theorem

Crooks' Fluctuation Theorem (CFT) [99] is an alternative approach to retrieving the free energy profile from a non-equilibrium process. Originally derived for non-equilibrium steady states [100], entropy production fluctuation theorems are a general group of statistical mechanical relations which are valid for far-from-equilibrium systems. The general form of the fluctuation theorem:

$$\frac{P\left(+\sigma\right)}{P\left(-\sigma\right)} \simeq e^{\tau\sigma} \tag{2.27}$$

relates the probability, $P(+\sigma)$, of observing an entropy production rate σ over a trajectory of time τ to the probability of observing the negative of that rate, $P(-\sigma)$. The fluctuation theorem expresses the probability of the entropy of a non-equilibrium system flowing in the opposite direction to that according to the second law of thermodynamics.

Crooks' Fluctuation Theorem relates the free energy difference between two states, ΔF, to the work done during the transition between the two states according to:

$$\frac{P_{0\to\Lambda}(W)}{P_{\Lambda\to 0}(W)} = e^{(\beta(W-\Delta F))} \tag{2.28}$$

Here, $P_{0\to\Lambda}(W)$ is the probability of expending the specified amount of work in the forward process, while $P_{\Lambda\to 0}(W)$ is the probability of expending the negative of the forward work in the reverse process.

CFT can be applied to a reversible system where the average work for the forward $\langle W_{0\to\Lambda}\rangle$ and reverse $\langle W_{\Lambda\to 0}\rangle$ process can be measured. The average work from repeated runs can then be related to the free energy by simply taking the average of the forward and reverse average work:

$$\Delta F_\lambda = \left(\langle W_{0\to\Lambda}\rangle_\lambda - \langle W_{\Lambda\to 0}\rangle_\lambda\right)/2 \tag{2.29}$$

The free energy difference can then be related to the PMF directly as before. CFT is advantageous due to the removal of sampling error introduced by estimating the average ensemble exponential of work in the Jarzynski equation; neither does it introduce truncation errors as a cumulant expansion is not required. The major disadvantage of course is that the process of concern must be reversible. If the process can be approximated to and hence performed in a reversible manner, then CFT should be employed to minimise errors.

CFT has been used to reconstruct the free energy profile from the forward and reverse processes of potassium ion translocation through the transmembrane protein Gramicidin A [101]. JE was also used to calculate the free energy profile of the process and CFT was found to be more accurate. The resulting free energy profile indicated different polarisation responses of the ion to the bulk water and the lipid membrane. The results

represent the successful use of CFT for the translocation of a small, symmetric molecule at speeds far from equilibrium (10 - 30 Å/ns).

In order to investigate if CFT could be used in the systems explored in this paper, we attempted to perform the reverse process of forward translocation of nucleic acid polymers through αHL. Forward pulling translocations were performed and snapshots of the system's atomic positions and velocities were taken. Reverse pulling simulations were then performed using these snapshots. Here, the SMD atom was pulled in the opposite direction to the forward process. Using this method at infinitesimally slow speeds or vanishingly small translocation distances, the reverse pulling will be the reverse process of the forward translocation. The reverse process may not mirror the forward process if the system is far from equilibrium.

By comparing the molecular configurations and the free energy profiles of the forward and reverse processes, and varying the pulling speed and translocation distance, we examined if any convergence towards the reverse process was observed. By reducing the translocation speed, a small degree of convergence was noted. However, even at our slowest pulling speed of 0.002 Å/ps, translocation is two or three orders of magnitude faster than that which occurs experimentally and the reverse process is significantly different to that of the forward process in terms of both the molecular configuration of the pulling molecule and the free energy profiles. By reducing the forward pulling distance before reverse pulling was attempted, while convergence to a reversible process was observed in the molecular configurations, very poor convergence was still observed in the free energy profiles. With such a large deformable translocating molecule, we found application of CFT under far from equilibrium conditions to be unfeasible.

2.5 Molecular Dynamics Simulations of Nucleotide Translocation through α-Hemolysin

By applying MD simulations to biomolecules in a pseudo natural environment, details of individual particle motions as a function of time can be represented,. This can uniquely address questions about the properties of a model system [102]. For instance, when applied to proteins, it is possible to understand their function, how they fold, or how they respond to other molecules. Such insight can provide the basis for drug design [103].

For the last 30 years, MD simulation studies of biomolecules have largely focused on the structure, folding, and interactions of proteins [104]. However, there has also been a considerable amount of research focused on MD simulations of DNA and RNA, including investigations into their structure, conformational changes, and interactions with proteins and ions [1].

α-Hemolysin is a well established and widely used MD simulation model. The crystal structure has been determined to a 1.9 Å resolution [38], which has allowed its use as a PDB file (atomic positions and sequence) in simulations. αHL has been analysed in detailed simulations on a lipid membrane and in a 1 M KCl water solution comprising a total of about 300,000 atoms to monitor fluctuations in the shape, behaviour under a transmembrane ionic current, the osmotic permeability and electro-osmotic effect, finding good agreement with experimental data [105].

With a working model of αHL, simulations have already given insight into the SCCR translocation process with respect to nucleic acid chains. Mathé *et al.* have discriminated $5'$ and $3'$ single-stranded DNA threading through αHL using experiments, but have also simulated the different translocations using a coarse-grained model [56]. They did this by creating a mathematical surface to represent the pore interior, then driving the translocation by a transmembrane potential [56]. As shown in Figure 2.5, Mathé *et al.* produced the translocation distance of the polymer with respect to time, showing a significant difference in the two orientations. They also provided a visual representation (shown in Figure 2.4) of the threading, which showed that the bases tend to point towards the $5'$ end. This nucleotide base orientation gives rise to a higher energetic barrier when translocation led by the $5'$ end occurs.

Another study using an approximate pore-model has given insight into experimentally observed current-trace patterns [106]. Using a protein model comprised of appropriately sized and charged beads in place of residues, it was observed that a commonly found stepping formation of current-traces may arise from the translocating DNA strand lingering in the vestibule of αHL before entering the transmembrane barrel. As is the benefit of MD simulations, a visual scheme of the process was produced to observe this effect, as shown in Figure 2.6.

The above examples were able to be performed on current supercomputing resources because coarse-grained models were used. By doing so, the number of calculations per

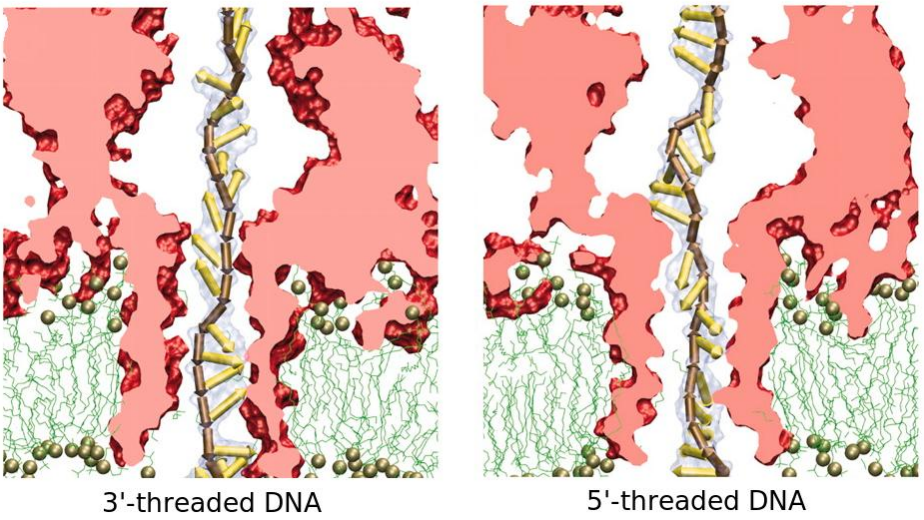


FIGURE 2.4: Difference in base orientation in 5′ and 3′ threaded ssDNA strands [56]. Note that the nucleotide bases tend to point towards the 5′ end of the chain in both cases. This creates a larger energetic barrier for 5′ threading as the bases must re-orient themselves to a larger degree in order to pass the constriction.

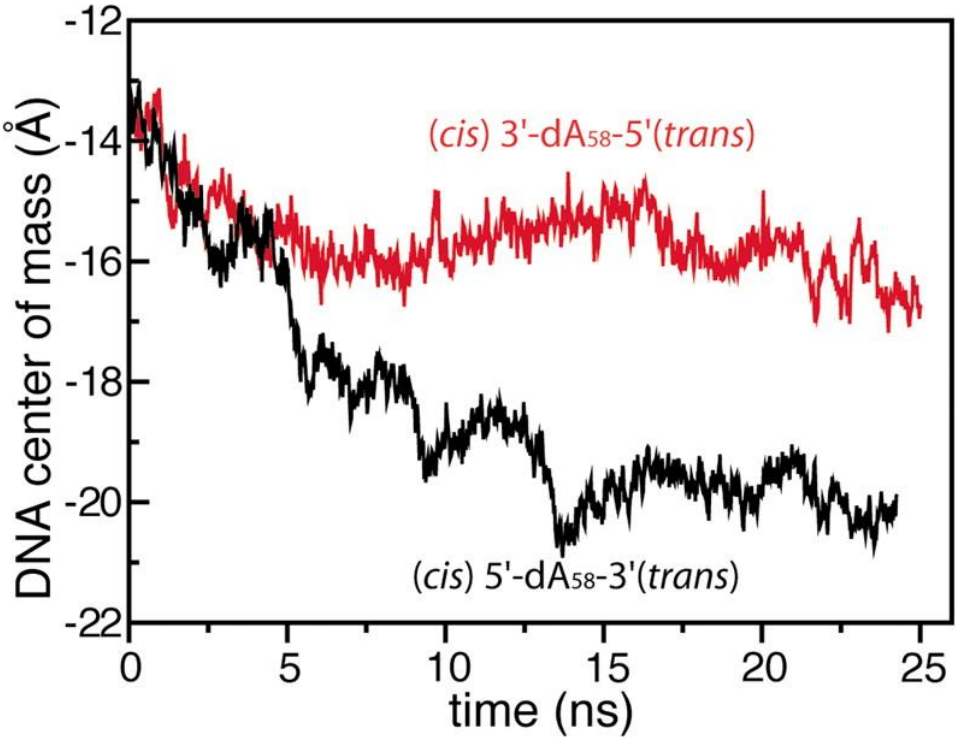


FIGURE 2.5: Translocation distance (Å) versus time (ns) for 5′ and 3′ threaded dA_{58} ssDNA strands [56]. The difference in translocation distance that arises due to the opposing orientations is significant.

timestep is reduced, which decreases the CPU power required and therefore the real time needed to run the simulation. By introducing approximate models, the simulations

are permitted to represent a longer timescale. The trade-off is that atomistic detail is lost, leading to a choice of balance in terms of quality versus quantity. These example could be performed in atomistic detail if the transmembrane potential was ramped up sufficiently. However, doing so can cause disruption of the system, for example the lipid membrane may break up [107].

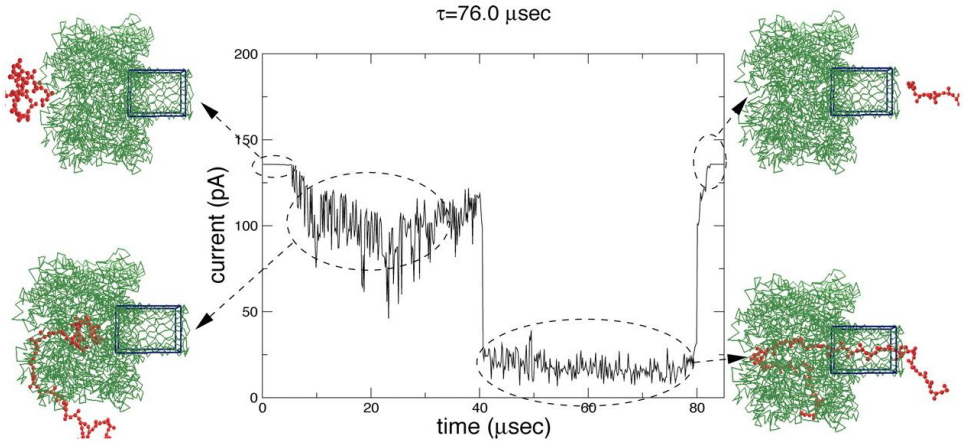


FIGURE 2.6: Simulated translocation of ssDNA with visual captures of the process with respect to time [106]. αHL is shown in green with the *cis*-entrance on the left, the *trans*-entrance on the right, the box highlights the transmembrane barrel, in red the nucleic acid chain is represented. The initial current blockage is shown to be due to the polymer entering the inner chamber of αHL.

A relatively new method of simulation translocation, called Grid-SMD, induces high speed translocation by applying a transmembrane potential that is weighted by a three-dimensional electrostatic potential map [107]. Grid-SMD allows an exaggerated transmembrane potential without disruption of the model, yielding successful translocations, and exhibiting polynucleotide conformations that are in line with expectations from experimentalists. As it is a new method, the limitations of using this method of translocation are currently ill-defined. Grid-SMD capabilities have been incorporated into the readily available MD simulation package, NAMD [92], although it has yet to be used for retrieving free energy profiles of translocation events. In the introductory study of the method, Grid-SMD was used for atomistically detailed models for the translocation of single stranded nucleic acid chains through αHL driven by an electric-field in tens of nanoseconds [107]. The simulations revealed a relationship between the separation of stacked nucleic acid bases at the αHL constriction and temporary halts in the translocation followed by surges.

2.6 Aims

Chapter 1 demonstrated the current relevance and the potential importance of the polynucleotide-nanopore translocation system. Chapter 2 has shown that, through computer simulations, it is possible to gain insight into such systems and produce important thermodynamic quantities such as the free energy of a process. Through a greater level of understanding unlocked by simulations, we can understand more about biological processes, and it is possible to design improved experimental components and conditions in, for instance, single channel current recordings.

Until the work in this thesis began, investigations of the αHL pore in polynucleotide translocation simulations were limited to coarse-grained models, lacking in atomistic detail. As the study by Wells *et al.* showed [107], it is possible to maintain atomistic detail by employing alternative means of translocation. This thesis investigates the translocation of polynucleotides through αHL nanopores, harnessing advanced methods of translocation, namely cv-SMD and ABF. Despite a broad uptake of both cv-SMD and ABF in MD simulations, a systematic comparison of the two methods has not yet been performed. The cv-SMD and ABF methods are investigated in Chapters 3 and 4 respectively, with a detailed comparison of the two at the end of Chapter 4.

Chapter 3

Constant Velocity-Steered Molecular Dynamics

The theory behind cv-SMD and JE, and the justification for its use in the polynucleotide translocation problem, has been discussed in Chapter 2. In this chapter, cv-SMD/JE is used to produce free energy profiles from translocation in various nucleotide-nanopore models. First, the method is presented, including the model construction, simulation parameters, computational resources and error calculations. Then, key optimisation simulations are outlined, followed by simulation results of single nucleotide and polynucleotide translocation through wild type and mutated αHL nanopores.

3.1 Method

This section reports the construction of the simulation model and the selection of software and parameters. Then the computational resources used and the calculation of errors are outlined.

3.1.1 Model Construction

The αHL crystallographic structure coordinates were taken from Protein Data Bank (PDB) entry 7AHL. Atoms missing from the PDB file (residues dLys-30, gLys-30, aLys-75, dLys-240, fLys-283, and aArg-66) were reconstructed using *psfgen*, which is part of

the NAMD2 [92] structure building module. αHL residues were mutated when required using the VMD [3] plug-in *mutate*. The protein was inserted into a patch of 150 Å x 150 Å pre-equilibrated and solvated 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid bilayer using the VMD plug-in *membrane*, aligned to the *xy*-plane plane. The centre of mass of the hydrophobic belt of αHL (residues 118-126 and 132-142) was aligned with the centre of mass of the lipid bilayer. Lipid and water molecules overlapping the protein were removed. The system was solvated in a water box of pre-equilibrated TIP3P [108] water molecules using the VMD plug-in *solvate* with a 2.4 Å buffer distance between water molecules applied. The aqueous solution was set at 1M NaCl using the VMD plug-in *ionize*. The dimensions of the lipid membrane and water box were chosen to minimise interactions between periodic images of the molecules, due to the use of three-dimensional periodic boundary conditions. Figure 3.1 shows αHL inserted in a lipid membrane as in our models.

The protonation states chosen were consistent with SCCR recording pH range, typically between 7.5 to 8.5 [37, 56]. The protonation states of the amino acid and phosphate groups were selected as follows: the amine groups of lysine and arginine residues were protonated, carrying a positive charge; inter-chain phosphate groups were unprotonated, carrying a single negative charge; finally, terminal phosphate groups on the single nucleotide molecules were unprotonated, carrying a double negative charge.

The poly(A) and poly(dC) molecules (molecular structures depicted in Figure 3.2) were constructed using the AMBER module *nucgen* [68]. Nucleic acid polymers with 25 bases were constructed; the length was selected in order to maximise the influence of the nucleic acid polymer helix throughout the simulation, allowing comparisons to experimentally used lengths of 100-200 bases while keeping the number of atoms in the model reasonably low. Double stranded conformations of the nucleic acid polymers were first built, the coordinates of which were derived from fibre diffraction studies [110]. A selected strand of the double helix was then removed from the PDB file, while the remaining single stranded nucleic acid polymer was charge neutralised with sodium counter-ions using the AMBER module *xleap* [68]. The nucleic acid polymer was then orientated as necessary using the VMD module *orient* [3] and the C3′-carbon atom of the leading residue was aligned with the center of the alpha carbon atoms (C_α) of protein residue 111. Water molecules that were overlapping the inserted polynucleotide were removed. After pulling the SMD atom from the constriction to the *trans*-entrance, the 25-base polynucleotide

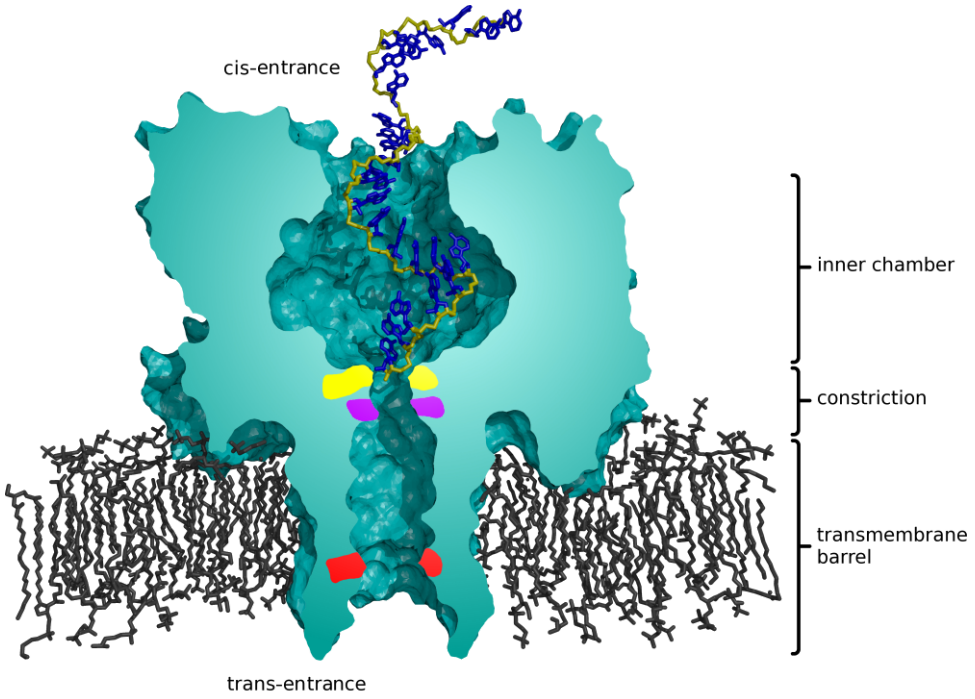


FIGURE 3.1: Figure representing the starting configuration of a 3′ led A_{25} translocation simulation. The heptameric protein pore αHL (green) is inserted into a lipid bilayer (black). Features of the translocating molecule include the backbone of A_{25} (dark yellow) and the nucleic acid bases (blue). The *trans*-entrance is at the bottom of the pore; taking the *trans*-entrance of αHL as a reference point at 0 Å, other notable features include protein residue Leu-135 at 13 Å (red), Met-113 at 43 Å (pink), Lys-147 at 45 Å (light yellow), and the *cis*-entrance at the top of the protein at 95 Å. The *cis*-entrance is 28 Å in diameter, the wide section of the pore running from the *cis*-entrance to residue Lys-147 is termed the inner chamber and is up to 46 Å wide. The constriction marked by residues Lys-147 and Met-113 is 14 Å wide, while the transmembrane barrel runs from the constriction to the *trans*-entrance and is around 20 Å wide; the *trans*-entrance is 24 Å wide. The C3′ carbon atom of the 3′ end residue of A_{25} is aligned with the center of mass of the C_α atoms of protein residue 111, which lies at the mouth of the constriction, just above residue Lys-147. References to the z-axis in this thesis refer to the axis running from the top to the bottom of this figure, along the length of the pore. For the sake of clarity, water molecules, sodium and chloride ions are not displayed (they are found along the entire length of the pore). Image prepared using VMD [3] and Raster3D [109].

will have more than 10 bases remaining in the wide inner chamber. Our simulations with a solvated polynucleotide show that 10 bases are sufficient for it to form a helical structure in poly(A). An example of the starting position of the polynucleotides is shown in Figure 3.1.

Single nucleotide PDB files of adenosine (A_1) and deoxycytidine (dC_1) monophosphates were obtained (PDB identifiers AMP and DCM respectively). The topology files were modified accordingly to produce accompanying PSF files. The structures of these molecules are depicted in Figure 3.2.

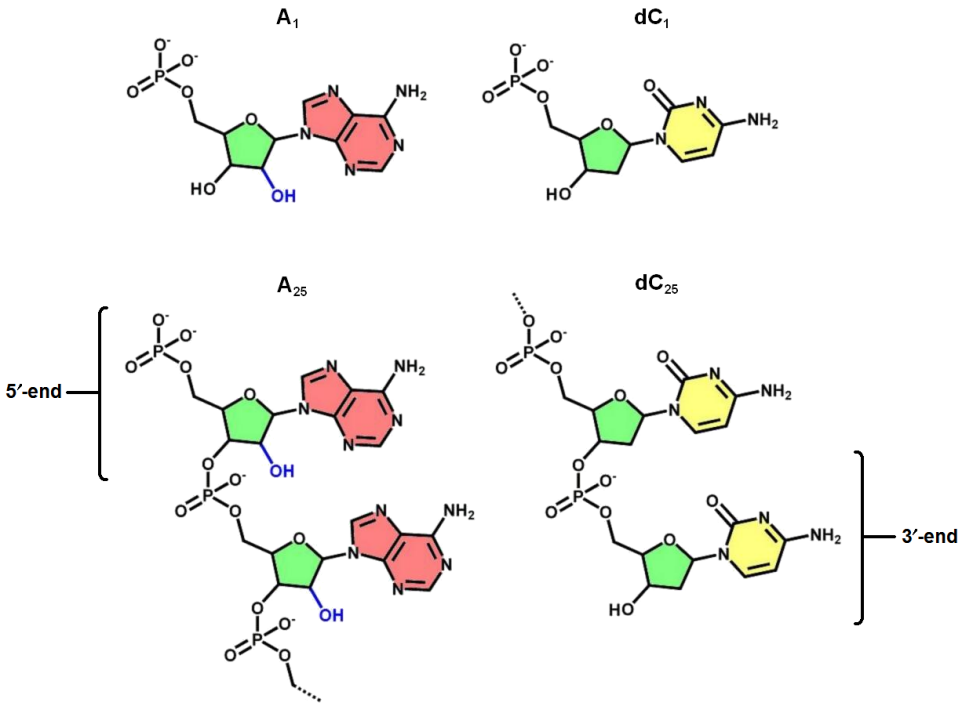


FIGURE 3.2: Molecular structures of A_{25}, A_1, dC_{25}, and dC_1. Structural features include the extra hydroxyl group associated with RNA (blue), the pentose base (green), the adenine base (light red) and the cytosine base (yellow). The 3′ and 5′-ends of the polynucleotides are also labelled for reference.

The final models consisted of 328,000 and 262,000 atoms for the 25-base polynucleotide and single nucleotide models respectively. The dimensions of the models were 145 Å x 145 Å x 154 Å for the 25 base nucleic acid polymer system and 145 Å x 145 Å x 121 Å for the single nucleotide system. The protein and nucleic acid molecules were at a minimum of 12 Å from the edge of the simulation cell, ensuring a minimum distance of 24 Å from the periodic images was maintained.

Translocation was chosen to be led by the 3′-end (see Figure 3.2) of the polynucleotide for all the simulations in this paper. As shown by several groups, a difference is observed in the translocation times of nucleic acid polymers depending on whether they are threaded by the 3′ or 5′ end of the nucleotide chain [56, 111, 112]. Nucleic acid bases projecting out along the phosphate backbone must flatten towards the backbone in order to pass through a tight constriction such as found in αHL. Nucleic acid bases tend to point towards the 5′ end; therefore, they must bend through a larger angle than for 3′-threaded polymers in order for the molecule to translocate through the pore. This results in an energetically less favourable process for 5′ translocation than for 3′-threaded instances. Thus 3′-led translocation was selected to compare poly(A) and poly(dC).

Given the nature of cv-SMD pulling, the stressing of bond angles such as that noted for the deformation in the helical shape of the nucleic acid polymers is likely to occur to some degree in the angle of the bases along the backbone. This influences the 5′-led polymers more than the 3′ ones as the bases could be made to invert orientation towards the 3′-end by the force of translocation.

Translocation simulations in this thesis were focused on SMD atom movement through the constriction and the transmembrane barrel of αHL (see Figure 3.1). The dimensions of this part of the pore present the greatest resistance to a translocating molecule [105]. The starting position of the 25-base polymer is shown in Figure 3.1; the SMD atom of the leading nucleic acid residue is aligned with the centre of the C_α atoms of the glutamic acid residue 111 at the top of the constriction. The dimensions of the inner chamber of αHL allow for the presence of a polynucleotide helix, whereas the smaller diameters of the constriction and transmembrane barrel do not. Hence, at these constricting sections of the pore, the helix of a translocating polynucleotide is required to unwind in order to pass through.

Constraints placed on the movement of protein amino acid residues are necessary in order to prevent the entire pore from being translocated due to the cv-SMD applied force. To ensure that specific interactions between amino acid residues and nucleic acid residues could take place, the side chains on the protein amino acids were left unconstrained to allow reorientation. Constraints of 0.5 N/m were therefore limited to the C_α atoms of the protein amino acid residues.

3.1.2 Simulation Details

Simulations were performed using the molecular dynamics simulation package NAMD version 2.6 [92]. NAMD is designed for scalable high-performance simulations on parallel machines. The CHARMM [67] forcefield was applied using all-hydrogen parameter files for CHARMM22 proteins and CHARMM27 lipids and nucleic acids. Periodic boundary conditions were applied and the Particle Mesh Ewald method [113] was used to calculate electrostatic forces with a dielectric constant $\epsilon_0 = 1$, computed over 128 Å x 128 Å x 128 Å grids. Van der Waals energies were calculated with a smooth 10 to 12 Å cut-off. Constant temperature was maintained at 295 K using Langevin forces [114]. Constant pressure was maintained by Nosé-Hoover Langevin piston pressure control [115]

at 1.01325 bar. A 2 fs timestep was used, with SHAKE turned on. So as to eliminate high-energy atom clashes, models were energy minimised for 2 ps until the gradient tolerance was below 20. The temperature was raised by 5 K every 50 fs for 3 ps to a final temperature of 295 K. Prior to insertion of the nucleic acid, the αHL model was equilibrated in the NpT ensemble for 1.3 ns with the backbone of the protein constrained, and for a further 4 ns without constraints. After insertion of the nucleic acid molecules, the model was equilibrated in the NpT ensemble for 4 ns with the α carbons (C_α) of the protein constrained and the ribose unit's C3′ atom of the leading residue of the nucleic acid molecule fixed. Translocation was limited to 1 ns per simulation due to resource constraints and to control binning errors. An overlap of 0.2 ns between sequential simulations was performed to enable removal of startup artifacts.

In order to sample the ensemble, multiple simulations of the nucleic acid molecule translocating past the same section of the pore were required. The initial configurations used to perform these translocation samples were obtained by capturing snapshots of the atomic positions and velocities, separated by 0.2 ns at equilibrium, with the SMD atom position fixed and the C_α protein atoms restrained.

Table 3.1 lists the key system configurations simulated using cv-SMD. The molecular structures of the translocating molecules are depicted in Figure 3.2.

TABLE 3.1: Table of the translocating molecules and αHL-pore types simulated using constant velocity-steered molecular dynamics. 'Wild Type' indicates αHL with no mutated residues, 'Mutant' indicates αHL mutant L147M.

Pulling Method	System Name	αHL Type	Nucleotide Base	Nucleotides	Samples Performed
cv-SMD	A_1-WT	Wild Type	Adenine	1	16
cv-SMD	A_{25}-WT	Wild Type	Adenine	25	16
cv-SMD	dC_1-WT	Wild Type	Deoxycytosine	1	16
cv-SMD	dC_{25}-WT	Wild Type	Deoxycytosine	25	16
cv-SMD	A_1-Mut	Mutant	Adenine	1	16
cv-SMD	A_{25}-Mut	Mutant	Adenine	25	16
cv-SMD	dC_1-Mut	Mutant	Deoxycytosine	1	16
cv-SMD	dC_{25}-Mut	Mutant	Deoxycytosine	25	16

3.1.3 Computational Resources

The computational determination of translocational free energy profiles involves multiple distinct phases. The first stage involves solving for different parameter values. Moreover, it is not clear initially how many samples need to be considered. This is dependent on the errors, both systematic and statistical. Once the "optimal" values have been determined, the solution becomes less of a parameter search problem and more a matter of aggregating resources so as to get high-throughput (at the optimal values), in order to obtain minimum errors and maximum samples in the lowest possible time. In general, the time-to-solution is arguably the most critical metric here. In previous work [116], we described how we use multiple resources to obtain the optimal values of the parameters to a first-approximation; we also showed how the ability to aggregate multiple high-end computational resources on a high performance computing grid, coupled with suitable amenable algorithmic techniques such as cv-SMD, can lead to a significant reduction in the time-to-solution. Although the benefits of utilising multiple computational resources are significant, some of the challenges and barriers encountered in the process have been discussed [117]; several of the same concerns are found in many applications [118].

In laboratory experiments, the time taken for a nucleic acid polymer to translocate through protein nanopores such as αHL has been shown to be in the region of hundreds of nanoseconds to tens of milliseconds [49]. We list the parallel scaling performance and efficiency data for an atomistically detailed model containing 328,000 atoms in Table 3.2. As can be inferred from the table, using 32 processors, a 1 ns simulation for a 328,000 atom model would take 40.5 and 65.5 hours on the supercomputing systems Queen Bee [119] and HPCx [120] respectively. Therefore, for high atom-count systems such as ours, it is not practically feasible to perform a set of simulations in the millisecond timescale with current supercomputing resources. The situation is hardly improved at higher processor counts.

When using cv-SMD, although the fundamental time-step at which the simulation proceeds is the same (typically 2 fs), the application of a force speeds-up the translocation process; hence for a given wall-clock time, non-equilibrium simulations can simulate equilibrium processes that would have taken orders of magnitude greater wall-clock time. For the simulations in this paper, a single translocation simulation from the top of the αHL constriction to the bottom of the transmembrane barrel requires 12 ns. Even using

cv-SMD, with 32 processors the simulation times are not suitable for these models and translocation lengths if they are to be properly sampled and explored. With 512 processors, a 1 ns simulation for a 328,000 atom model on both Queen Bee and HPCx takes about 5.5 hours to perform. By using cv-SMD and 256 processors per simulation, we have been able to perform all of the simulations in this paper in an acceptable wall-clock time.

The simulations performed to generate the results we report here required more than 800,000 CPU hours; it is important to note that these were required on high-end tightly-coupled machines, including those on HPCx [120], NGS [121], TeraGrid [122] and LONI [123]. A total of about 300,000 CPU hours were used for the 25-base translocation simulations at 0.004 Å/ps, while 150,000 CPU hours were used for all unmutated (wild type) αHL single nucleotide simulations. The remaining 350,000 CPU hours were used for other explorative simulations, translocation simulations for alternative lengths of nucleic acid polymers, attempts at reversibility, simulations exploring different spring constant and pulling speed parameters, and mutated αHL simulations.

TABLE 3.2: Parallel scaling performance and efficiency data. The performance is reported in terms of wall-clock time taken in seconds per timestep, and the efficiency relative to 32 processor simulations for a 328k atom model. Efficiency, E, at i number of processors is calculated as $E\left(N_i : N_{32}\right) = \left(t_{32} N_{32}\right) / \left(t_i N_i\right)$ where N_i is the number of processors, N_{32} is the reference number of processors, t_{32} is the wallclock/timestep at 32 processors, and t_i is the wallclock/timestep using i number of processors. Data is presented for both the LONI/TeraGrid Queen Bee and HPCx machines.

	Queen Bee		HPCx	
Processors	Wallclock/ timestep (sec)	Efficiency	Wallclock/ timestep (sec)	Efficiency
32	0.29	1.00	0.47	1.00
64	0.16	0.91	0.25	0.95
128	0.11	0.67	0.13	0.90
256	0.06	0.65	0.07	0.83
512	0.04	0.48	0.04	0.74

3.1.4 Accounting for Errors and Uncertainty

Just as in any experiment, multiple sources of error arise when using simulations to calculate free energies. In addition to the usual errors arising from numerical imprecision and approximations, there are errors due to the finite sampling used to compute ensemble

averages, and the choice of simulation parameters. Additional sources of error arise due to the finite approximations that are made in the expression used for computing the free energy, as well as from the binning and smoothening of data. Errors can never be completely eliminated; for a fixed computational effort, at best the errors can be controlled and their influence minimised. For this reason, in this work, we have taken considerable care to discern the effect of the different types of errors and we have planned our simulations accordingly.

Two sources of error are introduced specifically by the cv-SMD methodology. Firstly, by increasing the force constant of the harmonic spring connecting the SMD atom and constraint position, systematic errors due to deviations of the SMD atom from the reaction coordinate are reduced. However, a high spring constant also introduces statistical noise. An optimal balance between systematic and statistical error needs to be found. Secondly, the statistical error from applying JE, in its direct exponential form, is reduced by the use of the cumulant expansion formula, which introduces a systematic error due to truncation of the higher order terms. The systematic error can be controlled through this method, so as to be smaller than the statistical error from exponential averaging, thus providing a better overall estimate. These sources of error were discussed in greater detail in Section 2.4.1.1.

In determining the values of the work and energy averaged over multiple samples, there are two types of errors that must be accounted for. The first arises from the use of a coarse-grained value for the work. The mean value is referred to as the bin-value, and the error that arises from replacing the individual data points with the average taken over the bin, i.e., binning, is referred to as the binning error. The binning standard deviation, σ_{bin} was calculated as follows:

$$\sigma_{bin} = \sqrt{\frac{1}{N_d}\sum_{i=1}^{N_d}(x-\bar{x})^2} \tag{3.1}$$

Where N_d is the number of data points, x the value of data point i, and $\bar{x}$ the current mean of x. The binning error was then calculated from σ_{bin} as the standard error:

$$SE_{\bar{x}_{bin}} = \frac{\sigma_{bin}}{\sqrt{N_d}} \tag{3.2}$$

If the binning error calculation script was to calculate a bin average by holding all of data relevant to that bin in the analysing computer's memory before performing the calculation, the computer would likely struggle from holding too much data at once. To avoid this, the bin mean value and the associated binning error was calculated as a running average with each new data point. Performing the calculation in this way means that Bressel's correction ($N-1$ in place of N) was not employed, however, the impact of the correction is negligible due to the size of N_d. For global plots, each bin represents the local work/free energy value for that bin, combined with the cumulative work/free energy for all previous bins. Thus the local binning errors propagate from the start to the end of the reaction coordinate, forming a global binning error. Binning errors are therefore calculated according to the propagation of errors for each bin of a global plot:

$$SE_{\bar{x}_{newglob}} = \sqrt{\left(SE_{\bar{x}_{oldglob}}\right)^2 + \left(SE_{\bar{x}_{newbin}}\right)^2} \tag{3.3}$$

Where $SE_{\bar{x}_{oldglob}}$ is the total global binning error from the start of the reaction up to, but not including, the latest bin being calculated. $SE_{\bar{x}_{newbin}}$ is the binning error for the bin currently being calculated, and $SE_{\bar{x}_{newglob}}$ is the total global binning error, including the contributions of the new bin.

The second form of error arises from taking the average of the bin-value over multiple samples (sample-to-sample error); this error is statistical in nature and can be reduced with a greater number of samples. A challenge arises when the number of samples required to lower the statistical error is impractically large. The sample-to-sample error was calculated from the standard deviation:

$$\sigma_s = \sqrt{\frac{1}{N_s - 1}\sum_{i=1}^{N_s}(x-\bar{x})^2} \tag{3.4}$$

Where N_s is the number of samples. Bressel's correction is significant here as N_s ranges from four to sixteen. The sample-to-sample error was then calculated from σ_{sample} as follows:

$$SE_{\bar{x}_{sample}} = \frac{\sigma_{sample}}{\sqrt{N_s}} \tag{3.5}$$

The set of calculated binning and sampling errors were then be combined according to the propagation of errors. Firstly the binning errors from N_s samples were combined to form a total binning error, $SE_{\bar{x}_{bintotal}}$:

$$SE_{\bar{x}_{bintotal}} = \sqrt{(SE_{\bar{x}_{bin1}})^2 + (SE_{\bar{x}_{bin2}})^2 + ... + (SE_{\bar{x}_{binN}})^2} \tag{3.6}$$

The total binning error and the sample-to-sample error were combined to form a total error, $SE_{\bar{x}_{total}}$

$$SE_{\bar{x}_{total}} = \sqrt{(SE_{\bar{x}_{bintotal}})^2 + (SE_{\bar{x}_{sample}})^2} \tag{3.7}$$

Additionally, as we explore in Section 3.2.1, a high pulling speed introduces undesirable conformational changes. Thus, a balance must be found between performing a significant degree of translocation on a given resource and obtaining correct and accurate data.

3.2 Results

The results from simulations performed using models is presented. The first part of the section deals with simulations designed to tune key parameters, the ideal values of which are highly system-dependant, requiring thorough examination. We then take a look at notable polynucleotide configurations that occur once the system has been equilibrated. The majority of the section then explores the free energy profiles from polynucleotide and single nucleotide translocation through wild type and L147M mutated αHL protein-pores, with molecular analysis of these findings.

3.2.1 Parameter Optimisation

As described in Section 3.1.1, there are many parameters that need to be carefully selected for effective cv-SMD simulations. For example, the choices of spring constant and pulling speed are especially sensitive to the size and conformational behaviour of the pulled molecule, requiring explicit attention in the translocation of polymers. As demonstrated by Park and Khalili-Araghi [94], a stiff spring connecting the SMD atom

and the constraint position is required in order to use JE to extract the free energy correctly. The positions of the SMD atom and the constraint position may be significantly different with a low spring constant, and greater deviation from the reaction coordinate takes place. A stiff spring is therefore essential; however, the statistical noise in the force output is higher for a larger force constant, thus a compromise between these factors is needed.

Figure 3.3a provides data which we use to find the optimised value of the spring constant for the translocation of dC_{25} in wild type αHL. At a pulling speed of 0.04 Å/ps, spring constants 50, 100 and 200 kcal/mol are shown to approximate a stiff spring as the SMD atom closely follows the constraint position. 100 kcal/mol was chosen to ensure the spring was stiff without introducing excessive statistical noise. Figure 3.3(b) shows that a 100 kcal/mol spring constant exhibits similar behaviour at pulling speeds of 0.04 and 0.004 Å/ps. The sampling rate for the SMD force was every 40 fs at 0.04 Å/ps and every 400 fs at 0.004 Å/ps.

The statistical and systematic errors tend to increase as the SMD atom moves further from its starting position. Lower spring constants allow the SMD atom to lag behind the constraint position as the translocation length increases. As shown in Figure 3.3a, the position of the SMD atom for 1.4, 7 and 50 kcal/mol spring constants shows an increase in lag between the SMD atom and the constraint position at the end of 4 Å of translocation. Translocations with 100 and 200 kcal/mol spring constants exhibit a very small increase in lag at the end of 4 Å. This is avoided by segmenting the reaction coordinate into 1 ns (4 Å) sections and piecing them together to form the full reaction coordinate. By doing so, the constraint position is reset to the SMD atom every 4 Å, controlling the onset of errors.

As shown by Wells *et al.* [107], if the system is too far from equilibrium the translocation of a large deformable molecule in a cv-SMD simulation can result in conformational changes. This is due to translocation forces being applied to only a small part of a molecule. If the molecule moves too fast, there is insufficient time for relaxation forces to return the structure to its equilibrium state. To reduce this effect, the speed of cv-SMD simulations, particularly for large molecules, should be kept to a minimum. The challenge is to balance the desire for higher speeds with the need for correctness and accuracy. To illustrate the deviation from equilibrium configurations, Figure 3.4 shows

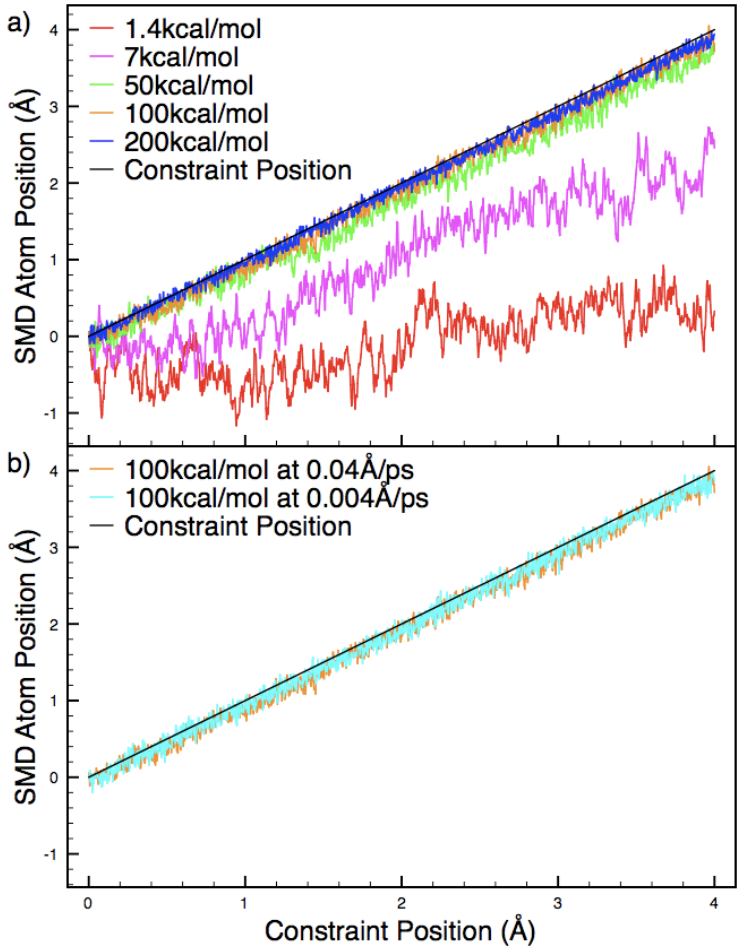


FIGURE 3.3: Analysis of the effect of spring constant values on the lag between the SMD atom and the constraint position. The profiles are calculated from single samples. (a) The position of the SMD atom and the constraint position (Å) plotted against the constraint position for simulations with differing spring constants. The simulations were run at 0.04 Å/ps using spring constants of 1.4, 7, 50, 100 and 200 kcal/mol. The harmonic spring must be approximately stiff in order for the free energy to be accurately calculated along a well defined reaction coordinate. This plot shows that the spring may be considered stiff at 100 kcal/mol, which was used for all subsequent simulations. (b) The SMD atom and constraint positions using a 100 kcal/mol spring constant pulled at 0.04 Å/ps or 0.004 Å/ps show that the spring is approximately stiff at both speeds, with only a marginal increase in stiffness at the slower speed.

the change in separation between the first and second, and the first and third, residues as the translocation of A_{25} proceeds at various speeds. At slower speeds, the backbone has more time to relax toward its equilibrium conformation, thereby keeping residues closer to each other during translocation. The increased separation is more pronounced as the residue under consideration is further along the chain from the pulled atom, highlighting the need to keep the pulling speed to a minimum. Taking these factors into account, a speed of 0.004 Å/ps was selected for the cv-SMD simulations reported in the remainder

of this chapter. A speed of 0.004 Å is roughly a thousand times faster than translocation speeds typically observed in SCCR experiments. While Akeson *et al.* [49] hypothesised that the unwinding of a poly(A) helix gives rise to a long translocation time, it should be noted that at a translocation speed of 0.004 Å/ps it is difficult to preserve the helical structure of the polynucleotide in the unconstricting inner chamber.

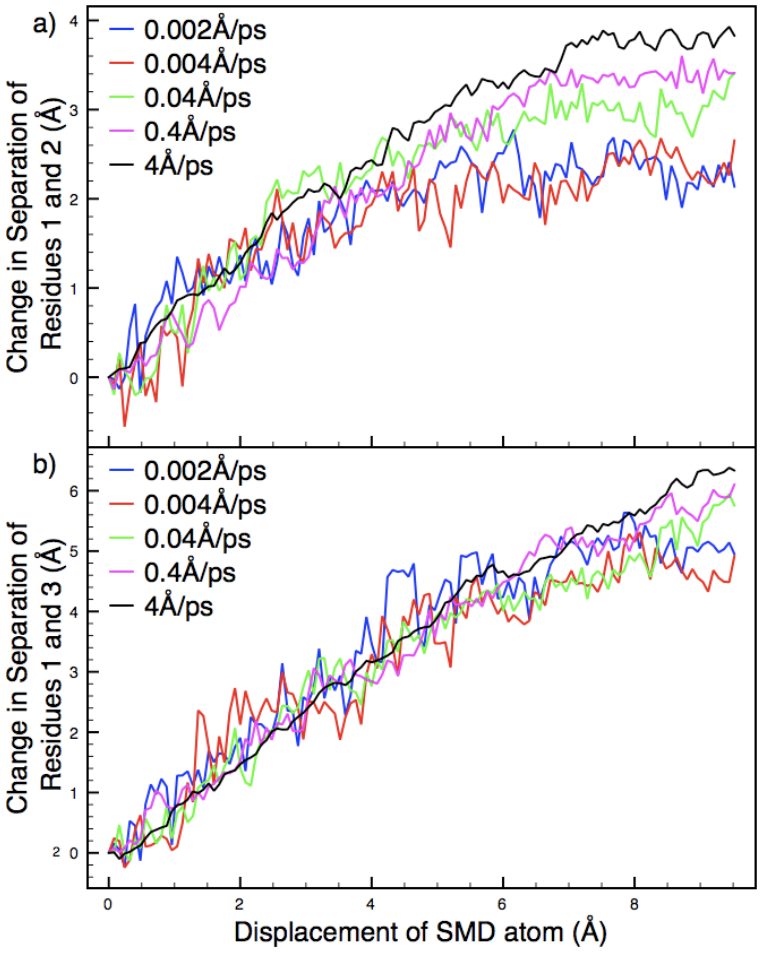


FIGURE 3.4: The effect of pulling speed on the relaxation of the conformation of translocating nucleic acid polymers. The profiles are calculated from single samples. The plots show the change in z-axis separation between the pulled atom (C3′-carbon atom of the front-most residue) and the C3′ atom of the (a) second and (b) third residues during an A_{25} pulling simulation at different speeds. At pulling speeds where the system is close to equilibrium, the separation between the first and other residues in the polymer will not tend to increase. Both plots show that at slower speeds, residues further along the polymer from the pulling atom have time to relax in conformation, and move closer to the pulled atom, reducing the stretching of the polymer. At higher pulling speeds the polymer stretches to a higher degree, making it desirable to keep the speed to a minimum. At a pulling speed of 0.002 Å/ps, the computational expense for our chosen translocation distances is too high; instead, 0.004 Å/s was chosen for simulations in this chapter.

3.2.2 Polynucleotide Simulations

In this section, simulations of polynucleotides in the αHL protein-pore are presented. In the first subsection, the equilibrium configuration of A_{25} and dC_{25} are compared. Then the free energy profiles from the translocation of these molecules are presented and analysed.

3.2.2.1 Base Stacking and Secondary Structure

In this subsection, conformations of A_{25} and dC_{25} at equilibrium are examined. Our simulations are consistent with the prediction that poly(A) exhibits a greater degree of base stacking than poly(dC) molecules [4–7]. This base stacking is thought to contribute to the 20-fold higher translocation time of poly(A) compared to poly(dC), observed in experiments [49]. Poly(A) possesses a helical structure with a diameter that is greater than the constriction and transmembrane barrel of αHL. According to Akeson *et al.* [49], the helix is believed to unwind during translocation, posing an energy barrier. It is known that overlap of the aromatic π-orbitals in the nucleic acid bases causes them to stack, giving rise to the helical conformation of the backbone. While the CHARMM27 forcefield does not explicitly describe π-orbital overlap, it accounts for the base stacking that results from the interactions between aromatic groups. The purine bases of poly(A) exhibit greater stacking than the pyrimidine bases of poly(dC) owing to the two aromatic rings and hence the greater π-orbital overlap of the purine groups (see Figure 3.2). The increased stacking confers on poly(A) a higher tendency to adopt a helical conformation in aqueous solution. Poly(dC) based polymers exhibit a more random structure due to a lower degree of stacking. On this basis, Akeson *et al.* proposed [49] that poly(dC) experiences a lower activation barrier to unwinding and so translocates through constricting pores faster.

At equilibrium, our systems containing poly(A) or poly(dC) exhibit a difference in mean conformation. As can be seen from Figure 3.5, which shows the average structures of A_{25} and dC_{25} in the inner chamber of αHL during a 2.6 ns simulation at equilibrium, A_{25} adopts a helical shape and the bases align in a well stacked formation. dC_{25}, on the other hand, displays a less defined helical backbone with reduced stacking of its bases.

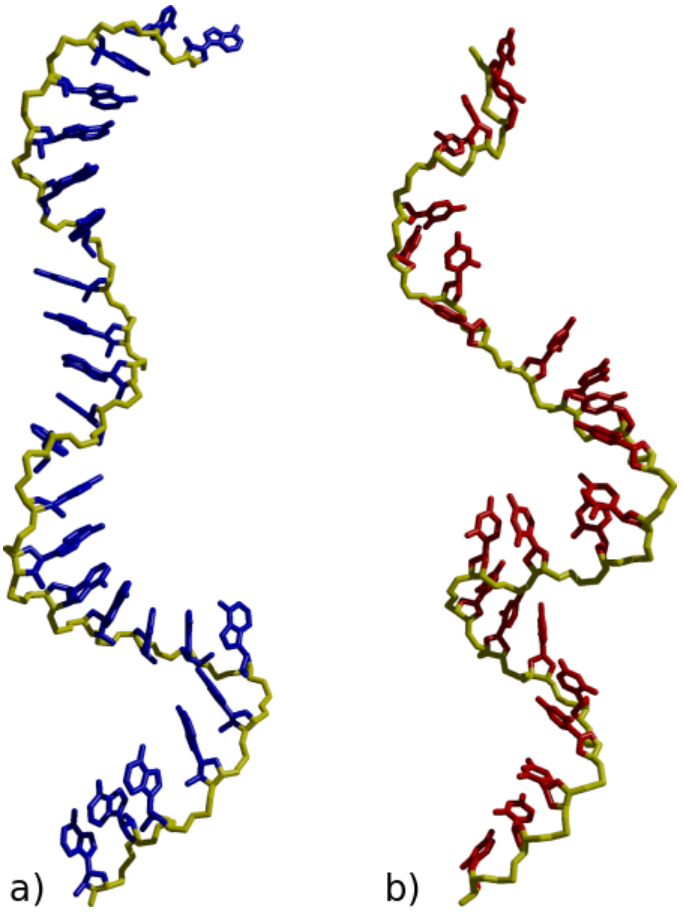


FIGURE 3.5: The average structures of (a) A_{25} and (b) dC_{25} over a 2.6 ns period at equilibrium in systems A_{25}-WT and dC_{25}-WT respectively (see Table 3.1). The C3′ atom of the front-most residue is fixed at the centre of the C_α atoms of protein residue 111 at the top of the αHL constriction. The polynucleotides lie in both the alpha chamber and outside the *cis*-entrance of αHL as shown in Figure 3.1. The A_{25} bases stack to a high degree, resulting in the more pronounced helical shape of its backbone than in dC_{25}. dC_{25} does exhibit base stacking, though this is significantly less than that of A_{25}, and is reflected in the more random configuration of the back bone. The increased stacking in A_{25} is due to the greater number of aromatic π-orbitals in adenine compared to cytosine, resulting in more favourable π-orbital overlap. Image prepared using VMD [3] and Raster3D [109].

As reported by Wells *et al.* [107], previous MD simulations have indicated that base stacking can influence the translocation time of nucleic acid polymers. These authors found that a break in the base stacking of a passing polynucleotide through the αHL constriction caused a temporary slowing down in translocation due to steric obstruction. Our own explorative simulations of stretched and unstretched polynucleotides indicate that base stacking can occur even in polynucleotides with a linear conformation. Therefore, energy barriers associated with the rupture of base stacking do not necessarily imply that helical unwinding has taken place. Base stacking and helical unwinding may

thus represent two energetic barriers to translocation.

While the average structures show a difference in base stacking and conformations of A_{25} and dC_{25}, we will find in the following sections that peaks in the local work and free energy profiles appear to correlate to interactions other than those relating to stacked or unstacked sections along A_{25} or dC_{25}.

3.2.2.2 Free Energy Profiles from Polynucleotide Translocation

In this subsection, the free energy profiles from the translocation of 25 base poly(A) and poly(dC) nucleotides (systems A_{25}-WT and dC_{25}-WT (see Table 3.1)) are presented. Here, cv-SMD is used to produce free energy profiles from the translocation of these polymers in order to determine whether the profiles reflect the experimentally determined longer translocation time for poly(A) compared to poly(dC) [49]. In addition, the simulations were conducted to gain physical insight into the differences.

The free energy profiles for a translocation length of 48 Å are shown in Figure 3.6 for A_{25} and dC_{25} with 16 samples performed for the calculation of each profile. Here, the SMD atom at the 3′-end of the nucleic acid polymer was pulled from the top of the constriction to the bottom of the transmembrane barrel. The pore dimensions, as listed in Figure 3.1, indicate that the steric barriers to translocation occur mainly within this region. Figure 3.7 represents snapshots of A_{25} polynucleotide conformations at 0, 12, 24, 26 and 48 Å of translocation during the type of cv-SMD simulation that produces the profiles found in Figure 3.6.

The global free energy plots for the 48 Å translocation of A_{25} and dC_{25} are shown in Figure 3.6(a). A_{25} displays a higher free energy profile than dC_{25}, with the discrimination beyond the error bars from 11 Å onwards. The separation between the means of the profiles continues to grow throughout the translocation process with the the free energy estimate for A_{25} being ∼30% higher than that of dC_{25} at the end of the 48 Å reaction coordinate. The higher free energy profile for A_{25} compared to dC_{25} is in qualitative agreement with the longer experimental translocation times for A_{25} [49].

Figure 3.6(b) shows the local free energy profiles of A_{25} and dC_{25}. Unlike in global profiles, local profiles error bars are dominated by binning errors. Therefore, in order to increase the confidence of each profile point, a larger bin width of 2 Å was used in the

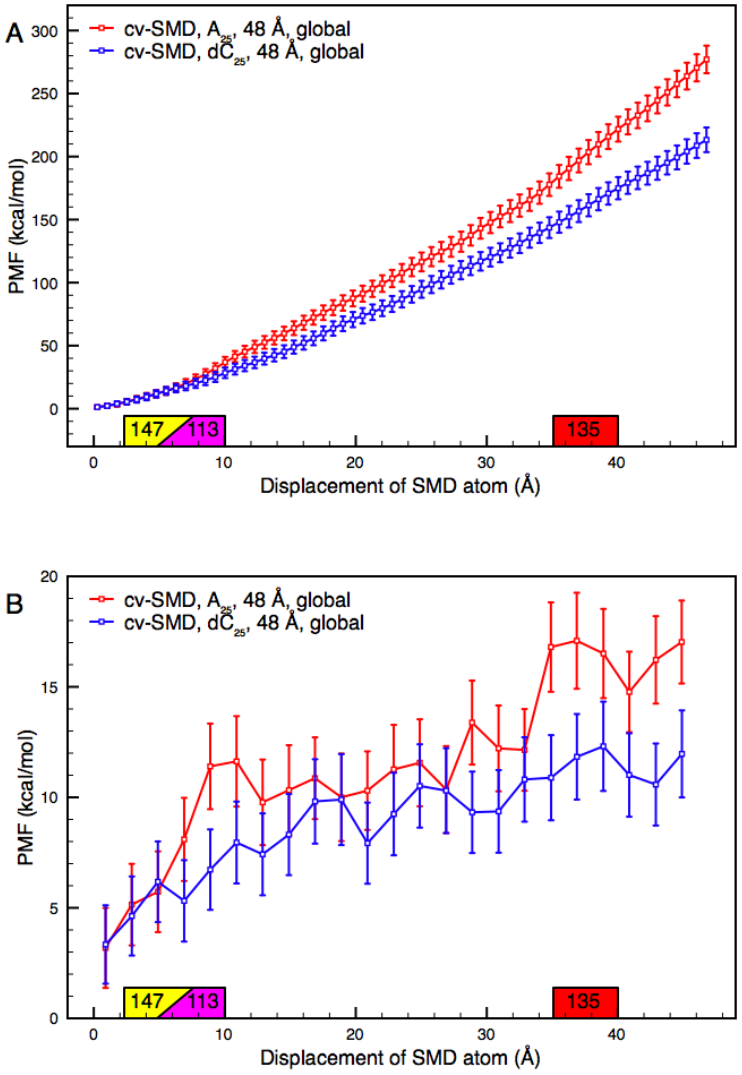


FIGURE 3.6: Local and global free energy profiles of A_{25} and dC_{25} translocation from the top of the constriction to the bottom of the *trans*-entrance of wild type αHL. A) Global free energy profile of A_{25} and dC_{25} translocation; each profile was derived from 16 samples, calculated using a bin width of 0.75 Å. Labelled along the x-axis are protein residues Met-147, Lys-113, and Leu-135. The residue labels span 5 Å from when the pulled atom to first phosphate atom passes the labelled residue. The free energy estimate for A_{25} is ~30% higher than that of dC_{25} at the end of the 48 Å reaction coordinate. The plots show discrimination of A_{25} and dC_{25} beyond the error bars after 11 Å of translocation. The gradients of both profiles gradually increase, which is in line with expectations that pulling additional nucleotides into the confining dimensions of the transmembrane barrel raises the energetic barriers to further translocation. B) Local free energy profiles of A_{25} and dC_{25} translocation; each profile was derived from 16 samples, calculated using a bin width of 2 Å. For these systems, the local environments lead to consistently higher energetic barriers to translocation for A_{25}. The A_{25} profiles also exhibits larger peaks than dC_{25}, most notably at 9 Å and 37 Å.

calculations. This increased the size of N_d dramatically, acting to reduce the size of the error bars at the cost of profile shape detail. Such local profiles act as useful complements

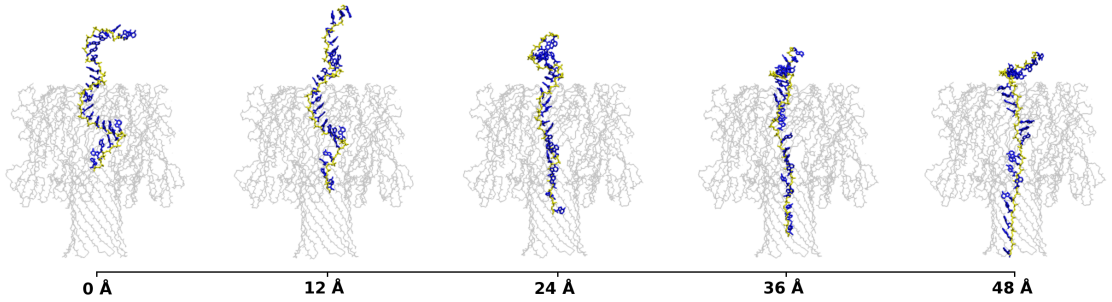


FIGURE 3.7: Molecular representations of A_{25} polynucleotide conformations at 0, 12, 24, 26 and 48 Å of translocation during a cv-SMD simulation. The images are single sample snapshots. The translocation was performed at a speed of 0.004 Å/ps. Only the polynucleotide backbone (yellow) and adenine bases (blue) are displayed. The backbone of the protein (grey) is shown for reference. The polynucleotide molecule stretches during the translocation event, giving a straight conformation in the beta barrel at 48 Å of pulling. Image prepared using VMD [3].

to the global free energy profiles to show peaks arising from localised contributions that contribute to the global profile. By examining the peaks in the local profiles, A_{25} can be seen to exhibit higher local barriers to translocation than dC_{25}. Large peaks found in the local work profiles of individual samples are considerably more pronounced than in the local free energy profiles consisting of multiple samples. The size of these peaks vary considerably from sample to sample, as do the peak's positions. Where there is some consistency in the peak positions, the local free energy profile exhibits peaks, notably at 9 and 37 Å for A_{25}, giving rise to A_{25} and dC_{25} separation beyond error bars. Key pore-residues near these positions are marked in Figure 3.6, the width of the markers represent where the leading nucleotide can access the protein residue in question. The single sample local work profiles also indicate that larger peaks occur for A_{25} than for dC_{25} (examples of single sample peaks are shown in Figure 3.9).

The molecular model of the system was examined in search of molecular reasoning for peaks appearing in the local work and free energy profiles. A_{25} was visually inspected at the pore constriction, finding that the phosphate groups of the polynucleotide strand interact electrostatically with the positively charged amine group of residue Lys-147. This interaction can be seen in Figure 3.8.

Figure 3.9 shows single sample local work profiles plotted against the radial distance of an interacting phosphate group and the most strongly interacting lysine residue. Similar patterns are seen in each of plots a) through d). Taking Figure 3.9(b) as an example, such patterns occur in the following manner: as the phosphate group associates with the

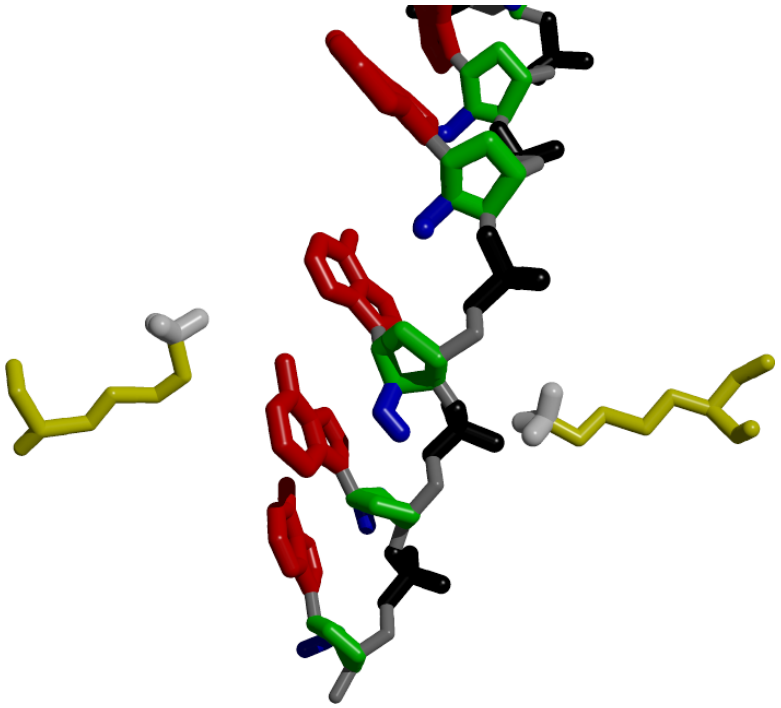


FIGURE 3.8: Molecular representation of the interaction between the nucleic acid phosphate groups of A_{25} and the amine group of residue Lys-147. The image is a single sample snapshot. The polynucleotide is shown in the middle with a lysine amino acid shown from either side of the constriction. The hydrogen atoms, other than those on the amine and hydroxyl groups have been omitted for clarity. Notable features include the polynucleotide's adenine group (red), pentose base (green), and the hydroxyl group (blue), which are coloured to match Figure 3.2. Also shown is the lysine residue (yellow), its amine group (white), and the polynucleotide's backbone phosphate groups (black). As a reference point, the Lys-147 residue is marked in light yellow on 3.1. The lysine group nearest the phosphates of the polynucleotide shows a reorientation of its side chain towards the phosphate while the lysine on the other side remains in a conformation that is observed during equilibration. Image prepared using VMD [3] and Raster3D [109].

lysine residue, a plateau is observed in the radial distance between the interacting groups. The radial distance fluctuates between 3.5 and 6.5 Å (right y-axis); in Figure 3.9(b), this plateau occurs between 38 and 43 Å of translocation (x-axis). The work profile (left y-axis) rises during this plateau in separation. This is because the polynucleotide backbone is being stretched between the interacting phosphate and the SMD atom. Then, as the attractive electrostatic interaction breaks, the radial separation of the phosphate and lysine increases. In Figure 3.9(b) this occurs between 43 and 44 Å. At the same time the work profile sharply drops from its maximum. There appears to be contributions from adjacent Lys-147 residues and adjacent phosphate groups along the chain, giving rise to the fluctuations of the plateaux between 3.5 and 6.5 Å during these times. The effect of adjacent phosphate groups on the work profile is shown in Figure 3.9(d). Here, the drop

from the work profile maximum is initiated by the formation of a second phosphate-lysine interaction, rather than by the breaking of the first.

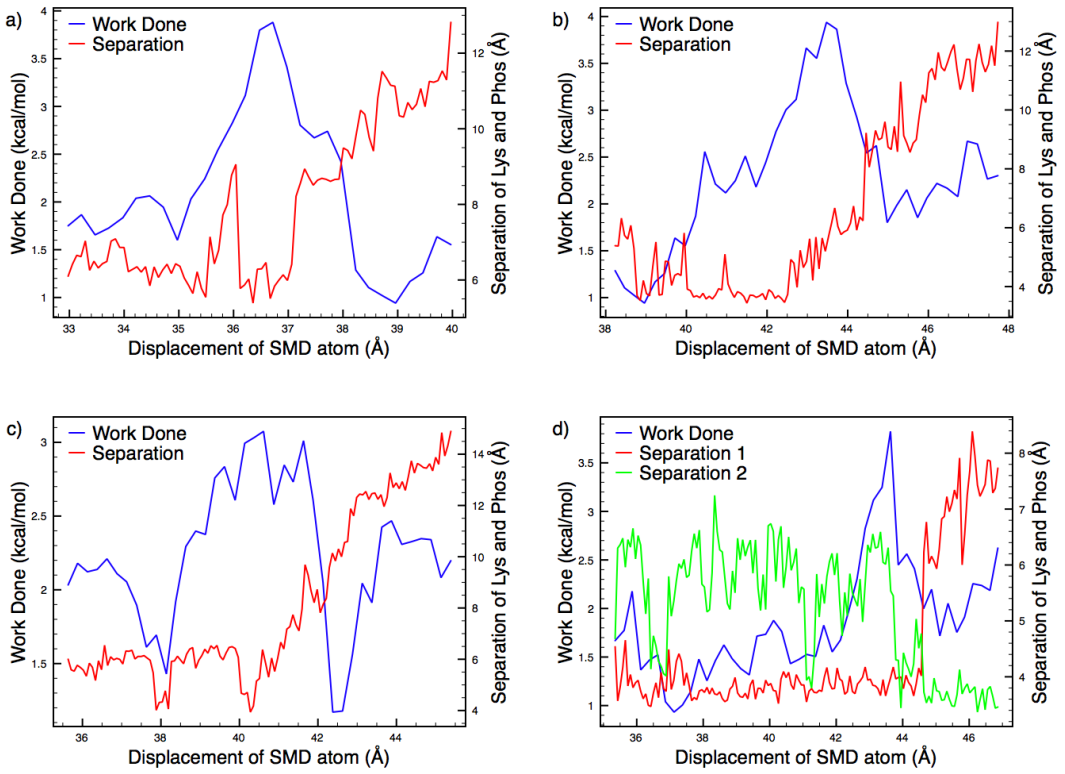


FIGURE 3.9: Selected peaks from local work profiles of A_{25} translocation, plotted alongside the separation between the phosphorus atom of a nucleic acid residue's phosphate group and the nitrogen on the Lys-147 side chain. The interactions plotted here were noted due to a significant re-orientation of Lys-147 side chains towards a passing nucleic acid phosphate group. The phosphate-lysine interaction is marked by a low plateau in the separation. (a), (b) and (c) A_{25} work profile plotted with the separation of the specified coinciding phosphate-lysine interaction. During the initial plateau in phosphate-lysine separation, the work profile can be seen to rise to a peak. As the interaction parts, marked by an increase in the separation of the lysine and phosphate group, the work profile sharply decreases. During a low plateau, the backbone between the interacting residue and the SMD atom tightens, causing the rise in the work profile. (d) A_{25} work profile plotted alongside the separation of the two phosphate groups (marked as Separation 1 and 2) interacting with a lysine group. Unlike with plots (a) (b) and (c), a notable difference is observed between the position of the work profile maximum and the increase in separation between the interacting phosphate group (1). Instead, in this instance the work profile sharply decreases when an adjacent phosphate (2) further along the polymer forms an electrostatic interaction with the same lysine group.

The local work and free energy profiles show higher peaks correlating to electrostatic phosphate-lysine interactions for A_{25} than for dC_{25}. This suggests a more accessible and/or a more electrostatically charged phosphate group on A_{25}. Base stacking or base size may play a role in making the A_{25} more accessible by pushing the phosphate group

close to the sides of the pore. Some of the molecular features of A_{25} (represented earlier in the chapter, Figure 3.2) may give rise to the different phosphate-lysine interaction compared to that of dC_{25}. The 2′ hydroxyl group on the A_{25} ribose sugar is separated from the phosphate group by three bonds and is radially very close. This hydroxyl group is inherent to RNA polymers such as A_{25}, and may influence the phosphate due to electrostatic as well as steric repulsion between the two groups. Figure 3.10 shows the influence that the RNA-hydroxyl group has on the angle between the phosphate group and the ribose sugar. The plot shows a greater angle where the phosphate group meets the ribose sugar in A_{25} than that of dC_{25}, indicating that an electrostatic/steric repulsion probably takes place between the two groups in A_{25}. The repulsion may affect both the electrostatic interactions exhibited by the phosphate group, and its steric accessibility due to the greater angle in relation to the ribose sugar.

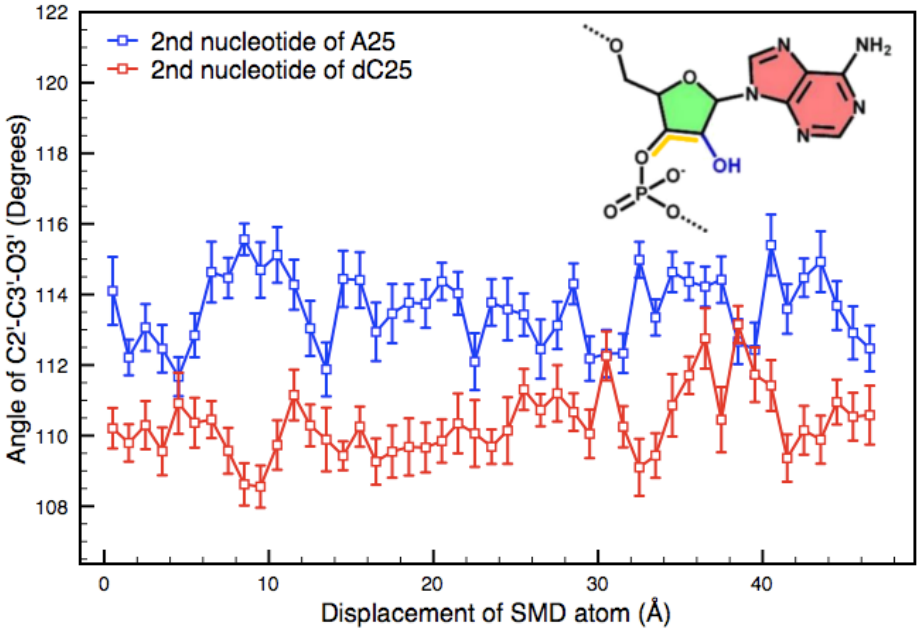


FIGURE 3.10: Plot of the angle between the phosphate group and the ribose sugar of polynucleotides during A_{25} and dC_{25} translocation simulations. The data is calculated from an average of two samples per profile, running the full length of the reaction coordinate. The error bars represent the sample-to-sample error. The plot illustrates the influence of the RNA-hydroxyl group on the phosphate group. The molecular structure of A_{25} is included to illustrate the angle investigated (yellow). The angle is between the C2′ and C3′ atoms of the ribose sugar and the O3′ atom of the attached phosphate group. The plot shows a higher angle in A_{25} than in dC_{25}, indicating that a repulsive electrostatic/steric interaction occurs between the hydroxyl and phosphate groups. The interaction may affect the electrostatics and steric accessibility of the phosphate group. These profiles show the phosphate-ribose angle for the second nucleotide in the chain, however the higher angle for A_{25} is consistently observed for all nucleotides.

If electrostatic interactions indeed play an important role in the translocation event, then it is worth noting other charged residues found in αHL. The constriction dimensions ensure that a translocating polymer will interact with lysine residue 147. However, there are other positively charged residues that may play a role, as shown in Figure 3.7. Lysine residues 21, 46, 51, 237, and 288 lie on top of αHL, residue 46 represents the highest point of the pore, and an incoming nucleic acid may interact with this before or during translocation. Lysine residue 8 protrudes into the pore at the *cis*-entrance; arginine residues 56 and 104, and lysine residue 154 are accessible on the lining of the inner chamber, while lysine residue 131 protrudes into the pore at the *trans*-entrance. In the simulations from this study, the charged residues of the inner chamber did not appear to interact with the polynucleotides to a significant degree, this could be in part because each residue was not as exposed as those on the edges of the constriction parts of the pore. Additionally, the backbones of the equilibrated polynucleotides tended not to approach the sides of the wide inner chamber. For the charged residues at the mouth of the *cis*-entrance and on top of the protein, the longest translocation events in this study only just began probe the pulling of the polynucleotide from lysine 8 at the *cis*-entrance at the end of the 48 Å reaction coordinate. Therefore, to investigate possible interactions with these additional residues would require a longer reaction coordinate, larger waterboxes accommodate the long distance-spanning polynucleotides, and therefore significantly more resources. It remains, however, a potential point of interest to return to in future studies.

3.2.3 Single Nucleotide Simulations

In this section, the free energy profiles from single nucleotide translocation in systems A_1-WT and dC_1-WT (see Table 3.1) are presented. The free energy profiles derived from polynucleotide translocation in the previous section represent combined energy barriers from the interactions of multiple nucleotides with the protein pore. To gain a clearer picture of the contributions from specific interactions, translocation simulations of single nucleotides were performed and the resulting free energy profiles are reported. By obtaining both the polymer and single nucleotide profiles, physical insight can be inferred with greater confidence. Single nucleotide translocation in protein-pores also possesses experimental relevance; by modifying the αHL pore with cyclodextrins and/or

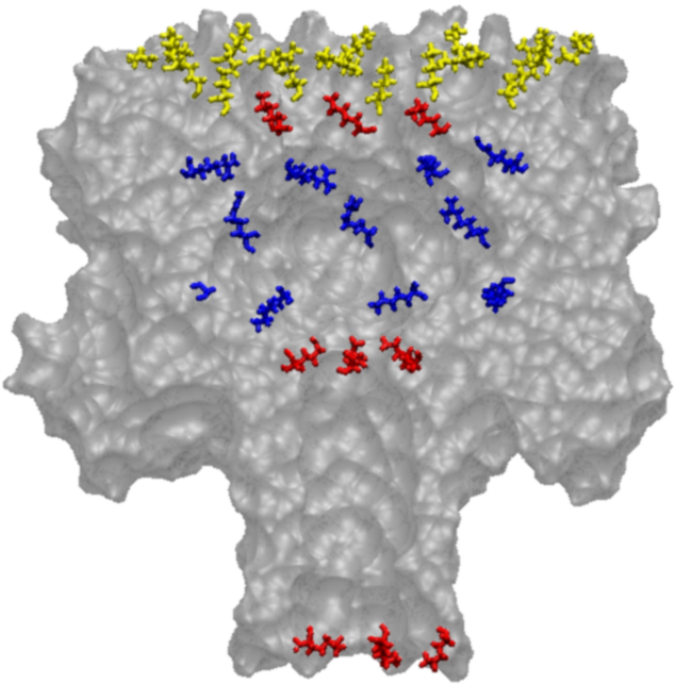


FIGURE 3.11: Positions of positively charged residues in αHL. A cross-section of αHL (grey) with positively charged residues that are accessible to a translocating molecule. These are lysine residues 8, 21, 46, 51, 131, 147, 154, 237, 288, and arginine residues 56, and 104. The residues are coloured to show that they are at a constriction (red), in the inner chamber (blue), or outside of the pore but still accessible to an incoming molecule (yellow). Image prepared using VMD [3].

exonuclease, it has been possible to distinguish different single nucleotide molecules in nanopore recording experiments [34, 124].

Using single nucleotide translocation simulations, it is relatively straight forward to locate the positions of key barriers to translocation. This is due to the ease of cross-referencing a peak in the free energy profile to a single nucleotide interaction. Figure 3.12 shows local free energy profiles A_1 and dC_1 throughout the pore constriction and the transmembrane barrel. This data represents two samples per profile, as this data is acting primarily as a probe into key energy barriers, additional samples are not necessary. The profiles show that the majority of activity occurs in the constriction region between 2 and 12 Å. There is a notable lack of activity outside this region, in stark contrast to the polynucleotide case, there is next to no activity at all around residue 135, which appeared to correlate to a peak in Figure 3.6(b). It can be inferred that this peak in the polynucleotide profile likely arose from an interaction further along the polynucleotide chain, away from the leading nucleotide.

Figure 3.12 also shows that, around the pore constriction, A_1 exhibits a broad peak between 2 and 12 Å, whereas dC_1 gives a large, sharp peak between 8 and 12 Å. Inspection of the molecular models around the sharp peak in dC_1 shows that the increase clearly correlates with a strong electrostatic interaction occurring between the phosphate group of dC_1 and the amine group of Lys-147. See our publication on this work [125] at http://pubs.acs.org/ for an animation of this process.

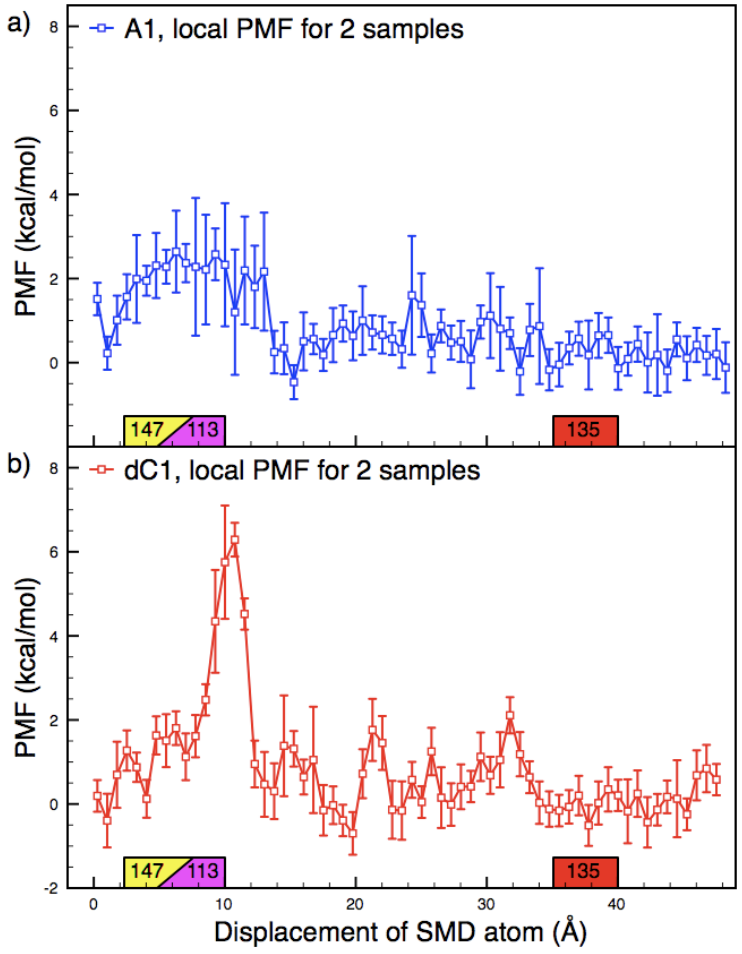


FIGURE 3.12: Local free energy profiles for A_1 and dC_1 translocation from the top of the constriction to the bottom of the transmembrane barrel; each profile is derived from two samples to probe the significant portions of the full reaction coordinate for single nucleotides. (a) Local free energy profile for A_1. A broad peak is shown between 2 and 12 Å which corresponds to the constriction of αHL. The error bars in this region are large, indicating a high degree of sample-to-sample variation. There is little notable activity exhibited in the transmembrane barrel. (b) Local free energy profile for dC_1. A sharp peak is observed at 10 Å and an examination of the model at this point reveals a strong interaction occurring between the phosphate group of dC_1 and the amine group of residue Lys-147 at the top of the constriction. The error bars in this region are small, indicating a low degree of sample-to-sample variation. Examination of this region in A_1 shows some degree of interaction between the phosphate group of A_1 and the amine group of residue Lys-147, though it is considerably reduced compared to dC_1.

With the most significant interval along the reaction coordinate identified, producing well sampled data for that region will allow advanced analysis to be reliable. Figure 3.13 shows highly sampled free energy profiles of single nucleotide translocation. Each profile was calculated from 16 samples. Figure 3.12 showed that there is little benefit, and great computational expense, in exploring single nucleotide translocation in the transmembrane barrel any further than 16 Å. Figure 3.14(b) gives the global free energy profiles, showing a modest ∼9% increase for A_1 compared to that of dC_1, with a large amount of overlap in the error bars. While the global profiles show little distinction, the local profiles given in Figure 3.13(b) reveal a key difference in profile shape. Here, A_1 exhibits a broad peak between 5 and 12 Å, while exhibits a sharper peak between 8 and 12Å. These peak widths are more visibly defined in the proceeding section when they are contrasted to translocation in mutated αHL pores.

Visual inspection of the single nucleotide molecules in the constriction gives insight into these different peak shapes. Comparing the molecular models of A_1 and dC_1 during equilibration and translocation we find notable differences in the local sodium ion positions. During the equilibration simulations of A_1 and dC_1 translocation systems, sodium ions can be seen to approach both molecules. A sodium ion approaches the A_1 molecule, associating itself between the phosphate and the purine base. The ion remains there for the rest of the equilibration and thus at the start of the translocation simulations. For dC_1, sodium ions approach the molecule but do not appear to associate with the pyrimidine base, moving away before the end of the simulation. During translocation, the A_1-associated sodium ion stays close to the phosphate group until the phosphate group moves past residue Lys-147; at this point the sodium ion remains above the constriction while A_1 translocates down the transmembrane barrel. See our publication on this work [125] at http://pubs.acs.org/ for an animation of this process. By contrast dC_1 only transiently receives an associated sodium ion during the translocation.

There are possible molecular reasons for the distinct behaviour of dC_1 and A_1 in relation to ions. Aromatic groups have a known tendency to form strong interactions with cations [126], just as we observe for A_1. Intuitively, one would expect the interacting sodium ion of A_1 to diminish the effective charge on the phosphate group and thereby reduce the interaction with the positively charged amine group of lysine. By contrast, it is anticipated that the uncoordinated phosphate of dC_1 will interact strongly with the

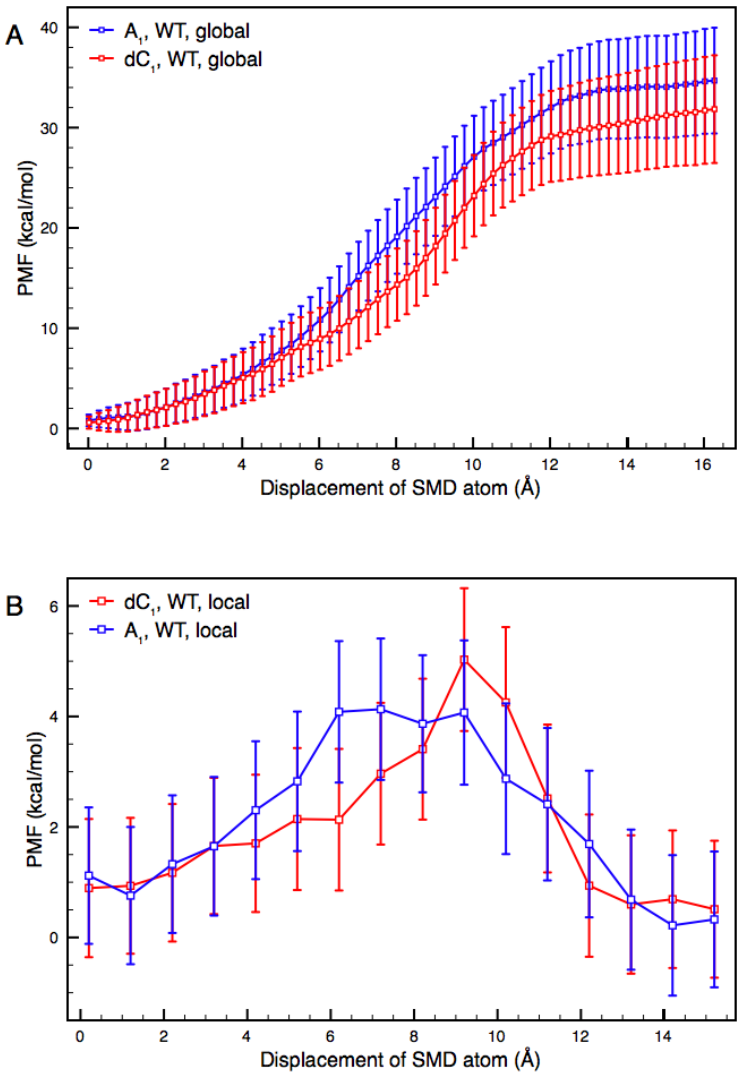


FIGURE 3.13: Global and local free energy profiles of A_1 and dC_1 translocation at the pore constriction for wild type αHL. A) Global free energy profiles for single nucleotide translocation; each profile was derived from 16 samples, calculated using a bin width of 0.25 Å. The ensemble average value for A_1 is ~9% higher than that of dC_1, and the error bars overlap considerably. B) Local free energy profiles for single nucleotide translocation; each profile was derived from 16 samples, calculated using a bin width of 1 Å. The profiles closely match each other than between 5 and 12 Å where the shapes differ, A_1 exhibits a broad peak compared to the sharp peak in the profile.

lysine side chain, as is apparent in the sharp, and repeatable peak in the energy profile in Figure 3.12a.

Figure 3.14 shows free energy profiles of polynucleotides A_{25} and dC_{25}, plotted against profiles of comparable single nucleotides. The plots show a clear distinction in overall shape; the single nucleotide profiles match the shape of the polynucleotide profiles, with slightly lower values, until about 10 Å along the reaction coordinate. Beyond this point,

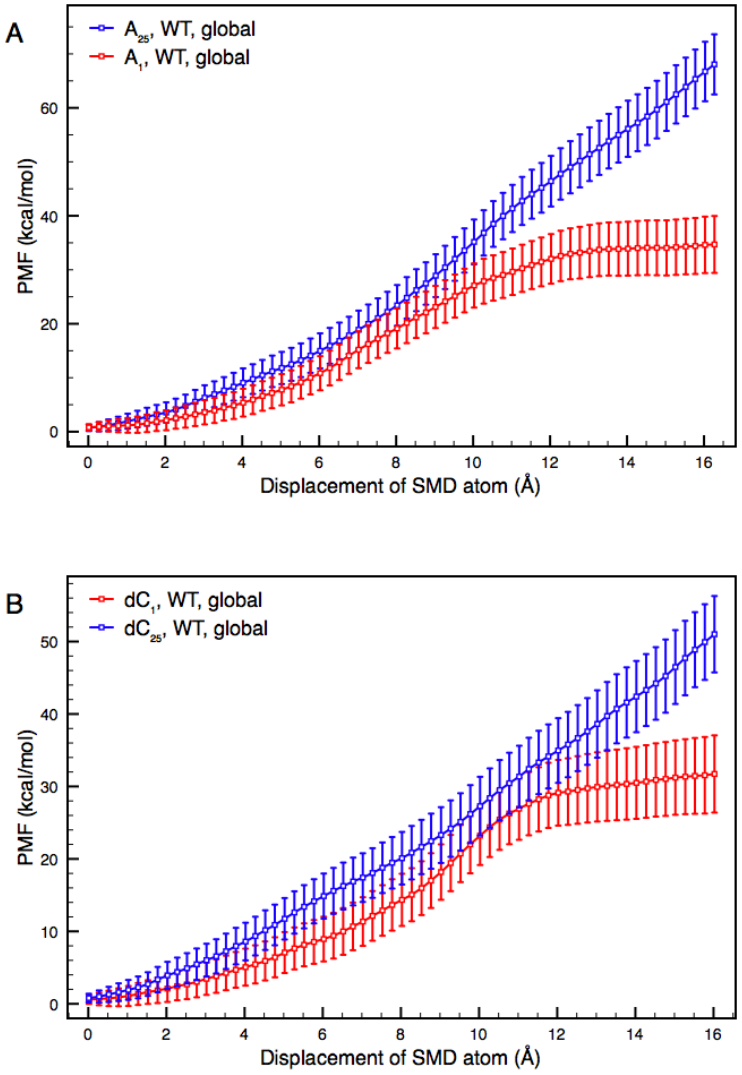


FIGURE 3.14: Global free energy profiles of A_1, dC_1, A_{25}, and dC_{25} translocation at the pore constriction for wild type αHL; each profile was derived from 16 samples. A) A_1 compared with A_{25}. The ensemble average values show separation beyond the reach of error bars after 10 Å of translocation. The profiles also exhibit distinct shapes; A_{25} shows a constant increase in free energy, while the A_1 profile levels off after 12Å of translocation. B) The ensemble average values show separation between the two profiles beyond the reach of error bars after 13 Å of translocation. The profiles also exhibit distinct shapes; A_{25} shows a constant increase in free energy, while the A_1 profile exhibits a steep gradient between 8 and 11 Å, then levels off after 11Å of translocation.

the polynucleotide profiles continue to increase while the single nucleotide profiles level off to near-zero gradient. This highlights the fact that the cv-SMD free energy is a calculation of energy barriers which affect the whole polynucleotide chain, rather than only experiencing the environment closest to the leading nucleotide.

3.2.4 Mutant α-Hemolysin Simulations

This section covers free energy profiles from polynucleotide and single nucleotide translocation in the dC_1-Mut and dC_{25}-Mut systems (see Table 3.1). From the single nucleotide and polynucleotide simulations in the unmutated αHL pore (see Table 3.1, "Wild Type" systems), the phosphate-lysine interaction appears to play a significant role. To confirm that this interaction contributes to the free energy profiles, simulations of translocation through a mutated αHL pore were performed, with the lysine at the constriction mutated into an neutral amino acid. Here, residue Lys-147 is mutated into methionine, as it is of similar size to lysine and uncharged. This mutation is not expected to alter the energetic barriers of the constriction beyond removal of the charge at protein position 147 at the constriction. This data also extends the investigation of cv-SMD for its ability to distinguishing different systems.

Figures 3.15 and 3.16 show the local and global free energy profiles from A_1 and dC_1 translocation through wild type αHL and L147M αHL . The global profiles of A_1 and dC_1 translocation (Figures 3.15(a) and 3.16(b)) show a higher free energy profile for translocation in the wild type pore compared the mutated pore, indicating a higher energetic barrier to translocation for both single nucleotides. The free energy values at the end of the reaction coordinate are $\sim$82% higher for A_1 in the wild type pore compared the mutant, and $\sim$53% higher for in the wild type compared to the mutant. Looking at the local profiles from these same simulations (Figures 3.15(b) and 3.16(b)), the distinction between the wild type and mutant pores becomes more clear. For A_1, the presence of the lysine residue induces a broad local free energy peak between 6 and 12 Å along the reaction coordinate. In dC_1, the presence of the lysine residue induces a sharp peak between 8 and 12 Å. These findings confirm the significance of the phosphate-lysine interaction.

Figure 3.17 shows the free energy profiles from polynucleotide translocation through wild type and L147M αHL pores. The difference between the wild type and mutated pores is expected to be reduced due to the phosphate groups in polynucleotide chains being singly charged, rather than doubly charged in the single nucleotide cases. In addition, intuitively the phosphate group is more accessible in single nucleotides than in polynucleotides. The figure confirms that the polynucleotides experience the phosphate-lysine interaction to a reduced degree, however, it also confirms that the interaction

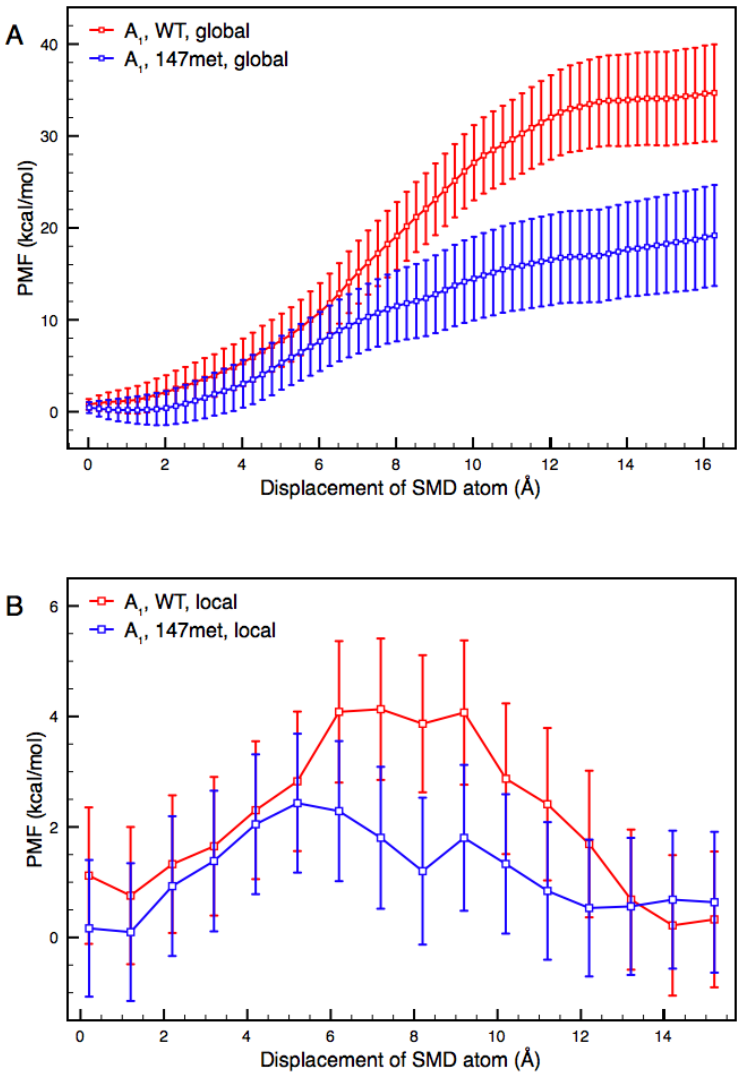


FIGURE 3.15: Global and local free energy profiles of A_1 translocation at the pore constriction for wild type and L147M mutant αHL pores. A) Global free energy profiles; each profile was derived from 16 samples, calculated using a bin width of 0.25 Å. The ensemble average value is ~82% higher for translocation in the wild type pore compared to that in the L147M mutant after 16 Å of translocation, the profiles separate beyond the error bars after 8 Å. B) Local free energy profiles; each profile was derived from 16 samples, calculated using a bin width of 1 Å. The profile shapes are markedly different; they closely match each other than between 6 and 12Å where a large peak indicates that the wild type pore exhibits a higher local barrier to translocation at this point.

does influence the translocation. In A_{25}, the free energy profile is ~17% higher for translocation in the wild type pore compared to that in the L147M mutant. In dC_{25}, the free energy profile is ~25% higher for translocation in the wild type pore compared to that in the L147M mutant. In both cases the profiles exhibit error bar overlap throughout, but tend towards separation, indicating that longer translocation would

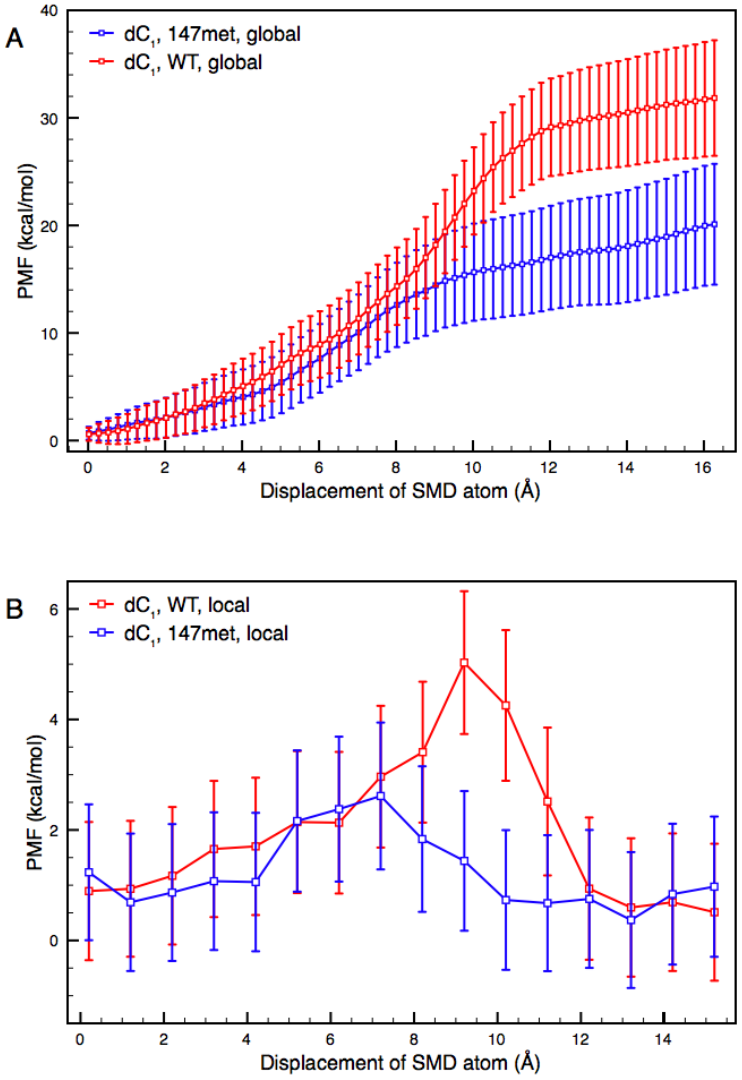


FIGURE 3.16: Global and local free energy profiles of A_{25} translocation at the pore constriction for wild type and L147M mutant αHL pores. A) Global free energy profiles; each profile was derived from 16 samples, calculated using a bin width of 0.25 Å. The ensemble average value is ~53% higher for translocation in the wild type pore compared to that in the L147M mutant after 16 Å of translocation, the profiles separate beyond the error bars after 10 Å. B) Local free energy profiles; each profile was derived from 16 samples, calculated using a bin width of 1 Å. The profile shapes are markedly different; they closely match each other than between 8 and 12 Å where a large and sharp peak indicates that the wild type pore exhibits a higher local barrier to translocation at this point.

lead to separation beyond error bars. The general trend of translocation in the wild type pore posing greater barriers to translocation remains visible.

The degree to which the phosphate-lysine interaction plays a role in single channel

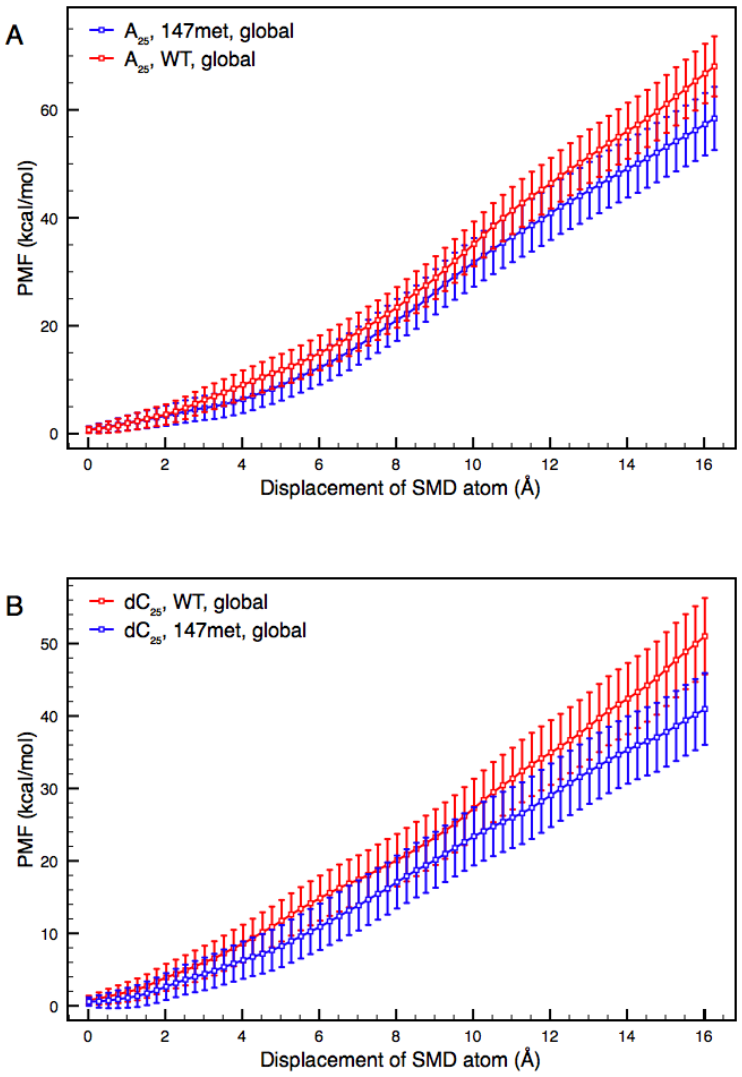


FIGURE 3.17: Global free energy profiles of A_{25} and dC_{25} translocation at the pore constriction for wild type and L147M mutant αHL pores; each profile was derived from 16 samples. A) A_{25} free energy profiles. The ensemble average value is $\sim$17% higher for translocation in the wild type pore compared to that in the L147M mutant after 16 Å of translocation. The profiles exhibit error bar overlap throughout, but tend towards separation, indicating that longer translocation would lead to separation beyond error bars. B) dC_{25} free energy profiles. The ensemble average value is $\sim$25% higher for translocation in the wild type pore compared to that in the L147M mutant after 16 Å of translocation. The profiles exhibit error bar overlap throughout, but tend towards separation, indicating that longer translocation would lead to separation beyond error bars.

current recording experiments is unknown. There are twelve positively charged protein residues accessible to a molecule translocating in αHL, and that current driven translocation is about one thousand times slower than in the simulations reported here. Conversely, under conditions pertaining in experimental translocation experiments there

exists a rapid ionic current flow, not present in our simulations, which may alter the influence of the electrostatic interaction. In nanopore recording experiments it has been found that increasing the number of positive charges inside the αHL pore increases the capturing frequency of polydeoxyribonucleotide translocations [127], indicating an electrostatic relationship between positively charged residues and negatively charged translocating polynucleotides. Our simulations show a high sensitivity to local ion positions, giving rise to greater sample-to-sample variance in the free energy profiles. Therefore, in order to draw comparisons between non-current driven translocation simulations and SCCR experiments, careful consideration of the role played by solvation-ions and counter-ions is required.

By focusing on the L147M mutated αHL pore simulations, it is possible to gauge further insight into contributions to the poly(A) and poly(dC) discrimination. Figure 3.18 shows global free energy profiles for A_1, dC_1, A_{25}, and dC_{25} translocation through the L147M mutated αHL pore. In Figure 3.18(a), A_1 and dC_1 exhibit remarkably similar free energy profiles and error bars. The profiles are almost indistinguishable throughout the reaction coordinate. This is interesting as it gives us a unique perspective on the influence that the increased sterics of adenine has compared to cytosine. Base stacking, helical unwinding, and charged interactions have all been eliminated from this data, leaving only the sizes of the bases as providing different barriers to translocation. The result is clear, the sterics of the base has no discernible impact on translocation. But now when we examine a comparable case for polynucleotide in Figure 3.18(b), we can see that discrimination between A_{25} and dC_{25} occurs without the presence of a charged amino acid at the constriction. Thus with base sterics and electrostatic interaction ruled out as causes, it can be hypothesised that base stacking between the adenine bases, and/or helical unwinding, is giving rise to this observed difference. This is in line with experimental expectations [49].

3.2.5 Animation descriptions

Two animations were described in Section 3.2.3. Here the animation features and details are described. They can be found at http://pubs.acs.org/

The first animation (phos-lys-interaction2.gif) represents the phosphate-lysine interaction during dC_1 translocation. Here, the grey molecules represent the backbones of

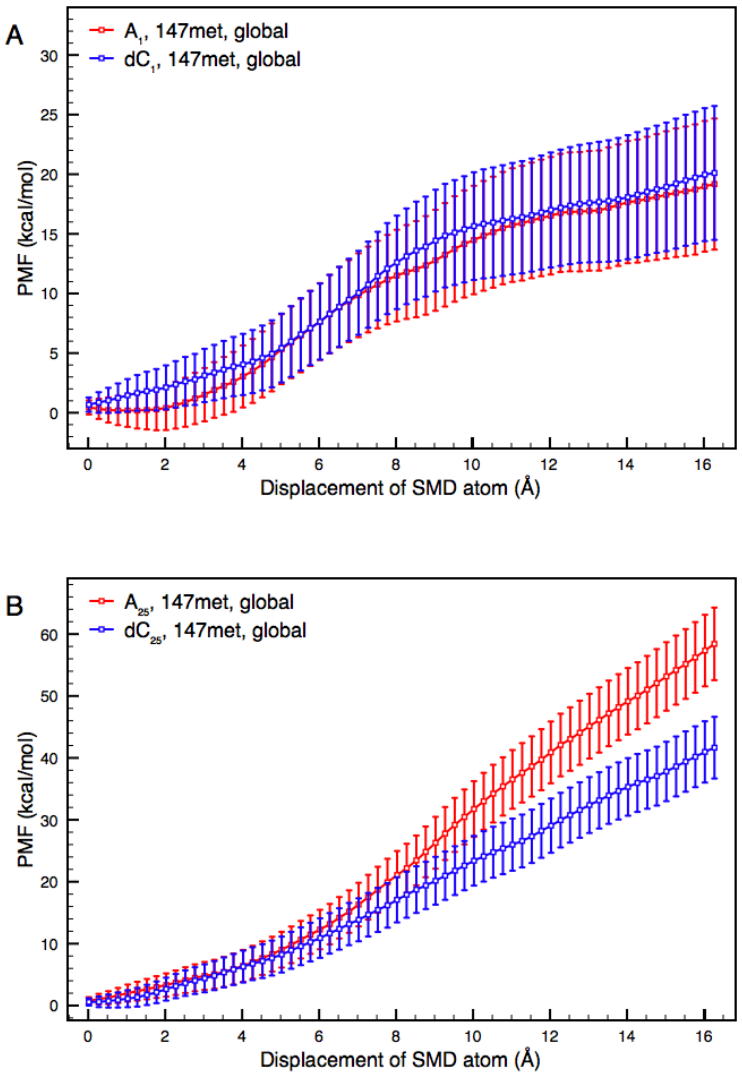


FIGURE 3.18: Global free energy profiles of A_1, dC_1, A_{25}, and dC_{25} translocation at the pore constriction for the αHL mutant L147M; each profile was derived from 16 samples. A) A_1 and dC_1 free energy profiles. The profiles show almost no discernible difference between them with near identical error bars and ensemble average values. B) A_{25} and dC_{25} free energy profiles. The ensemble average value for A_{25} is ~40% higher than that of dC_{25}. The profiles separate beyond the reach of error bars after 10.5 Å.

residues Met-113 and Lys-147 in the constriction of αHL. Two of the Lys-147 residues are coloured and their side chains are represented, in the middle is the translocating dC_1 molecule. Note the movement of the lysine side chains and the reorientation of dC_1 due to the attractive forces between the phosphate and amine group.

The second animation (a1-ion.gif) shows the A_1-associated ion during translocation at residues Met-113 and Lys-147. Note the cradled formation of the adenine and phosphate

group at the start and the sheering off of the ion as the nucleotide molecule passes Lys-147.

3.2.6 Constant Velocity-Steered Molecular Dynamics Conclusions

Using constant velocity-steered molecular dynamics in conjunction with Jarzynski's Equality, free energy profiles were produced from translocation simulations. These simulations covered single nucleotides and polynucleotides translocation through wild type and mutated αHL protein-nanopores, presented in Chapter 3.

Free energy profiles from A_{25} and dC_{25} translocation through wild type αHL exhibited the experimentally observed trend that A_{25} experiences greater barriers to translocation than dC_{25}. The free energy profiles of A_{25} was shown to be 30 % higher than that of dC_{25} after 48 Å of translocation. The local work and free energy profiles highlighted a key interaction that contributes to the translocation energy barriers, namely a phosphate-lysine interaction. We attributed higher peaks appearing in the A_{25} case to the influence of the RNA hydroxyl group.

Free energy profiles from single nucleotide translocation were used to clarify contributions to the polynucleotide free energy profiles. It was found that the barriers to translocation occur at primarily at the pore constriction. The constriction interactions were found to be different for A_1 compared to dC_1, with A_1 producing a wide peak in the free energy profile, and dC_1 producing a sharp peak. We attribute the difference to an interaction between the aromatic groups of adenine and a positively charged sodium ion, which also interacted with the negatively charged phosphate group. This reduced the phosphate-lysine interaction for A_1 relative to that of dC_1.

The lysine residue leading to the phosphate-lysine interaction was mutated for a sterically similar but uncharged methionine residue in order to decipher the impact on the free energy profiles. The distinction between translocation through the wild type and mutated pore was shown to be significant. By comparing the translocation of A_1, dC_1, A_{25}, and dC_{25} through the mutated pore, the near-identical single nucleotide profiles and readily distinguishable polynucleotide profiles gave strong evidence that the sterics of the adenine base compared to the smaller cytosine base play a negligible role in the barriers to translocation, pointing to base stacking as a likely cause of distinction.

Chapter 4

Adaptive Biasing Force

This chapter explores the use of the adaptive biasing force (ABF) methodology to produce free energy profiles from translocation in various nucleotide-nanopore models. The means of translocation in cv-SMD simulations is rigidly implemented; the atom position is moved at constant velocity, regardless of the free energy landscape. Through this mode of movement, the sampling of high energy components of the free energy landscape is allowed, together with the thermodynamic states of interest that they separate. The ABF methodology is another such technique that permits improved sampling of these thermodynamic states. In contrast to cv-SMD, ABF uses an estimation of the free energy landscape to calculate the biasing force. As discussed in Chapter 2, the biasing force in the ABF method is calculated during the simulation, requiring no prior knowledge of the free energy landscape. The theory behind ABF as a free energy calculation method has been established in Subsection 2.3.4.

By investigating ABF in addition to cv-SMD, the assessment of both methods can be improved by drawing critical comparisons between the two. This is important, as neither method has previously been used for an atomistically detailed model of the polynucleotide-nanopore translocation system, and so it is instructive to compare the different modes of translocation. The use of ABF presents a number of advantages compared to using cv-SMD. Firstly, ABF does not require the introduction of errors from facets of the cv-SMD/JE methodology (covered in Chapter 2). Also, ABF does not constrain the biased/SMD atom in axes orthogonal to the reaction coordinate, allowing full exploration of the internal pore dimensions. It is also worth noting that ABF may be

used in many more studies than cv-SMD as the reaction coordinate may be a measure of quantities such as bond angles, RMSD values and so on. We shall return to the comparison of ABF and cv-SMD at the end of the chapter.

Investigation of ABF in the nucleotide-nanopore system is explored in this chapter. The first section covers the methodological details such as the molecular model and simulation parameters. Then the optimisation and exploration of key ABF parameters is presented. Next the accounting of errors and uncertainty is discussed. Results of multi-sampled free energy profiles are then presented; these are produced from polynucleotide and single nucleotide translocation through wild type αHL. Finally the chapter draws direct comparisons of ABF to cv-SMD by simulating polynucleotide translocation with parameters set to be equivalent where possible, with parameters specific to the ABF or cv-SMD methodologies set to yield approximately equivalent conditions.

4.1 Method

Many of the parameters chosen for the cv-SMD simulations in the previous chapter represent optimal values for any MD simulation of the αHL -nucleotide system. Therefore, unless determined to be have unique consequences for either ABF or cv-SMD, the parameters were largely kept the same across both methodologies. The αHL protein-pore, nucleotide, lipid, solvent and ion models used for ABF simulations were identical to those used in the cv-SMD simulations described in Section 3.1.1 and shown in Figure 3.1. A biasing force was applied to the C3′-atom of the leading residue. NAMD2.7b1 was used for all ABF simulations, with the adaptive biasing force implemented using the *colvar* module [128]. Force measurements were accumulated in bins of 0.25 Å (unless otherwise stated) for 16 Å length trajectories. The reaction coordinate in the ABF methodology is calculated as a function of distance from the translocating molecule to a reference atom or set of atoms, in contrast to cv-SMD, this allows the protein to be left unconstrained. Unless otherwise stated, the simulations in this chapter were performed with an unconstrained αHL pore. The reaction coordinate was chosen to be the z-axis separation between the C3′-atom of the leading nucleotide and the centre of mass of the alpha carbon atoms of protein residue 111 - at the top of the constriction (see Figure 3.1 for a description of the orientation of the z-axis). While only the z-axis separation was controlled by the biasing force, the steric constraints of the pore interior were sufficient

to keep each sample trajectory within the desired *xy*-boundaries. The biased atom was kept within the outer *z*-axis boundaries of the reaction coordinate by a harmonic force implemented at either end. The length of the reaction coordinate was set to 16 Å, which the cv-SMD results indicated was sufficient for analysis of polynucleotide-nanopore interactions, and for producing separation in the free energy profiles of A_{25} and dC_{25} while keeping computational resource costs to a reasonable level. It is possible to split longer reaction coordinate into segments and reconstruct the full free energy profiles from each of the segments, this can help prevent the biased atom getting stuck. However, this is not necessary with a reaction coordinate length as short as 16 Å. Furthermore, introducing too many segments can cause the harmonic restraints at the ends of each segment to significantly impact the free energy values, and so it is best avoided. The number of simulated timesteps required to sample the reaction coordinate depends on the force measurement threshold parameter and the diffusion time, which varies between simulations. The simulations were therefore performed in blocks of 100,000 to 1 million molecular dynamics integration timesteps until the full reaction coordinate was sampled. As in the cv-SMD simulations, 1 timestep was equivalent to 2 femtoseconds. The width of the bins along the reaction coordinate and the force measurements threshold parameter (ζ) are investigated in the next section.

Table 4.1 lists the key system configurations simulated using ABF. The molecular structures of the translocating molecules are depicted in Figure 3.2.

TABLE 4.1: Table of the translocating molecules and αHL-pore types simulated using the adaptive biasing force method. 'Wild Type' indicates αHL with no mutated residues, 'Mutant' indicates αHL mutant L147M.

Pulling Method	System Name	αHL Type	Nucleotide Base	Nucleotides	Samples Performed
ABF	A_{25}-ABF-ζ5k	Wild Type	Adenine	25	4
ABF	A_{25}-ABF-ζ20k	Wild Type	Adenine	25	4
ABF	A_{25}-ABF-ζ80k	Wild Type	Adenine	25	1
ABF	dC_{25}-ABF-ζ5k	Wild Type	Deoxycytosine	25	4
ABF	dC_{25}-ABF-ζ20k	Wild Type	Deoxycytosine	25	4
ABF	dC_{25}-ABF-ζ80k	Wild Type	Deoxycytosine	25	1
ABF	A_1-ABF-ζ5k	Wild Type	Adenine	1	4
ABF	dC_1-ABF-ζ5k	Wild Type	Deoxycytosine	1	4

As was the case with the cv-SMD simulations in the previous chapter, these computationally intensity of these ABF simulations necessitates the need for high performance

computing facilities. The computational requirements for producing ABF data of comparable quality to cv-SMD data are roughly equivalent. A more detailed breakdown of this is presented towards the end of the chapter.

4.2 Investigation of ABF Parameters

In this section we explore the effects of key ABF parameters on the simulation results. The parameters in question are the force measurement threshold parameter and the bin width. The optimal value of these parameters is highly dependent on the system being studied and the computational resources available. Since the ABF methodology has previously not been tested on such a large translocating molecule in atomistic detail, it is important to investigate these parameters thoroughly.

4.2.1 Force Measurement Threshold Parameter

In ABF simulations the biasing potential is calculated in response to free energy barriers along the reaction coordinate. Therefore, unlike in cv-SMD simulations, there is no adjustable parameter that provides direct control over progression (movement of the biased atom from bin to subsequent bin) along the reaction coordinate. Once the system has been chosen, the force measurement threshold parameter, ζ, exerts the most control over progression along the reaction coordinate. ζ is a parameter, chosen in units of timesteps, which sets the number of instantaneous force measurements (which is equal to the number of timesteps) that are taken on the biased atom within a bin before the biasing force is applied. At least several hundred timesteps are required in order to make the average meaningful, due to the high variation in the instantaneous force. The length of time spent in the bin before the biasing force is applied also allows the system to relax from non-equilibrium effects introduced by the biasing force in preceding bins. In this section we explore the influence that ζ has on progression along the reaction coordinate and the value and form of the free energy profiles obtained from the simulations.

4.2.1.1 Influence on Trajectory Progression

Choosing a value of ζ has a direct effect on the time that the biased atom can spend in a bin along the reaction coordinate before a biasing force is applied. A higher ζ ensures a longer simulation and a slower average translocation speed. The nature of ABF means that the instantaneous translocation speed is variable, thus the average speed across the full reaction coordinate may be considered as an indicator of the impact of non-equilibrium effects. The average translocation speed is ideally kept to a minimum to ensure that non-equilibrium effects, and their impact on the free energy estimate, are kept low. A low average translocation speed can also lead to overly long simulation times, therefore a balance between speed and computational expense is required within a given computational resource budget.

The impact that the ζ parameter has on the biased atom's progression along the reaction coordinate is shown in Figure 4.1. The data is produced from simulations of polynucleotide (dC_{25}) translocation across a 16 Å trajectory in αHL. The figure shows the displacement of the biased atom plotted against the number of timesteps simulated at ζ values of 0, 5000, 20000 and 80000. At ζ=0 the translocation of the biased atom is extremely rapid; at ζ=5000 about 2 million timesteps are required in order for the biased atom to translocate the full length of the reaction coordinate; at ζ=20000 about 5 million timesteps are required; at ζ=80000 about 16 million timesteps are required. The figure indicates that the time taken for the biased atom to traverse the reaction coordinate shows a strong dependance on the value of ζ, though it is not directly proportional, the figure show that increasing ζ 4-fold results in less than a 4-fold increase in translocation time. This fact that the increase in translocation time is not directly proportional to the value of ζ is expected, since the value of ζ only directly affects the parts of the simulation when the biased atom does not yet have a biasing force applied to it. The value of ζ may influence the value of the applied bias due to its impact on reducing non-equilibrium effects, but this effect is unlikely to be directly proportional to the ζ value.

Figure 4.1 also shows that the fluctuations in reaction coordinate progression increase with a higher value of ζ. The degree of fluctuation in translocation progression is similar when ζ=5000 and ζ=20000. At ζ=80000, the translocation progression fluctuates to a significant degree. This relation of degree of fluctuation as a result of ζ is clear even

when the x-axis is normalised so that the x-axis end points are the same for each value of ζ (not shown). It should be noted that high fluctuations in the translocation progression can be viewed as evidence that the biasing force is accurately opposing the free energy barriers. This is because the ABF methodology does not aim to exceed barriers to translocation, but rather match them in order for self-diffusion to dictate translocation. In this way, translocation is permitted, rather than induced, meaning that with barriers effectively removed, translocation is likely to occur in both directions along the reaction coordinate. Only if these energy barriers are exceeded would translocation be expected to be unidirectional. The influence that fluctuating translocation progression has on the free energy profiles is shown in proceeding subsections.

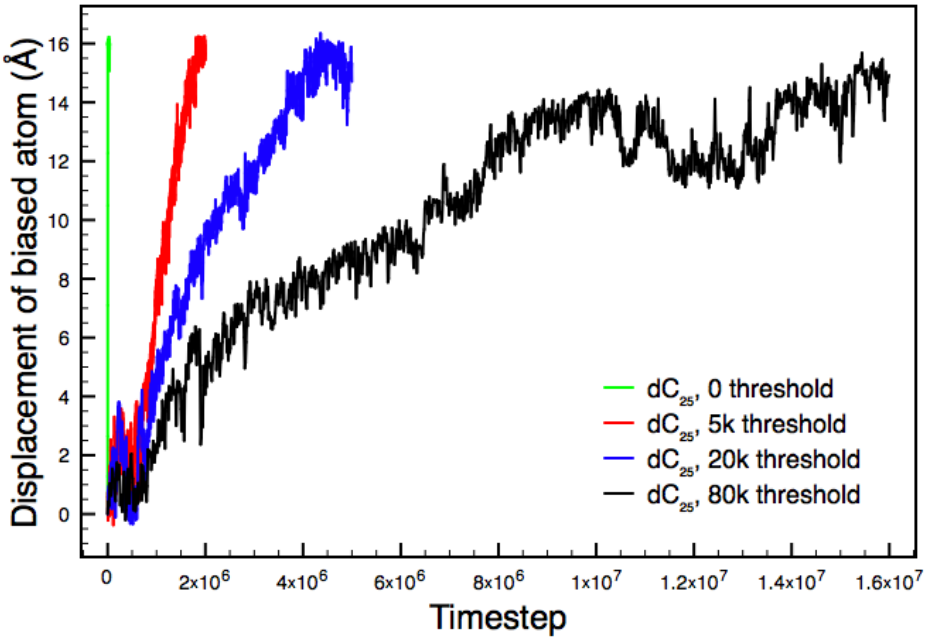


FIGURE 4.1: The z-axis displacement of the biased atom plotted against the simulation timestep in dC_{25} ABF simulations with ζ values of 0, 5000, 20000 and 80000 ζ. The time taken to traverse the reaction coordinate shows a strong dependance on the value of ζ. The tendency for the biased atom to move forward is seen to become less consistent at ζ=80000. The sample-to-sample variation in translocation time increased with higher ζ values, since a higher ζ value results in a greater dependance of the translocation on self-diffusion. The single samples represented here are sufficient to illustrate the effect of increasing ζ.

The size of the translocating molecule has a significant influence on its progression along the reaction coordinate. Figure 4.2 shows the displacement of the biased atom plotted against the number of timesteps simulated at ζ=80000 for dC_1 and dC_{25} translocation simulations. The progression along the reaction coordinate is seen to fluctuate to a

greater degree in the case of dC_1, resulting in a roughly 25% longer time to traverse the entire reaction coordinate. This trend can be explained on a molecular level; the smaller dC_1 molecule possesses greater freedom to diffuse in both directions along the reaction coordinate due to the size of the molecule. The biased atom of the polynucleotide molecule also has the whole chain to contend with, if it is to move in the reverse direction.

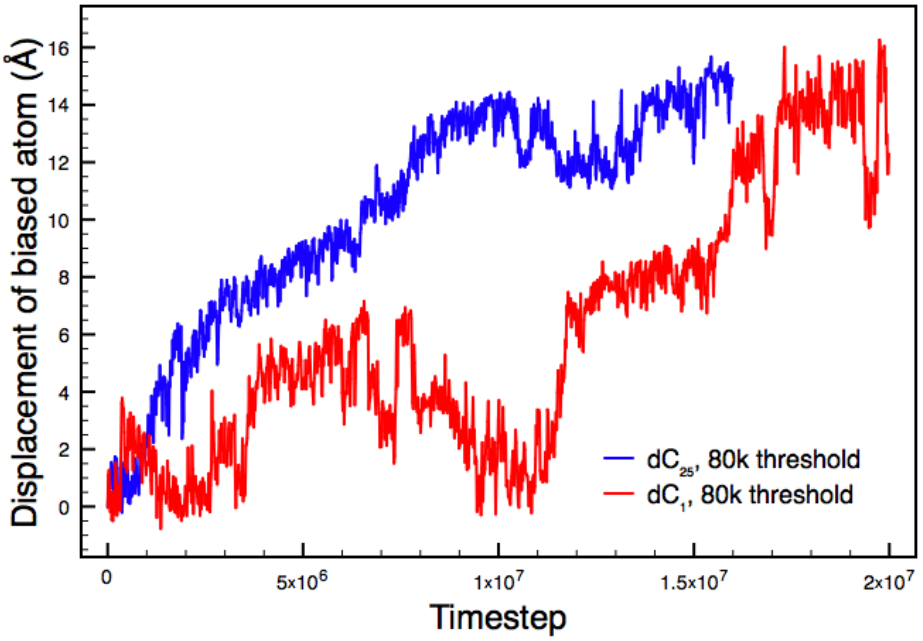


FIGURE 4.2: The z-axis displacement of the biased atom plotted against the simulation timestep in dC_{25} and dC_1 ABF simulations with a ζ values of 80000. The polynucleotide is shown to have a much greater tendency to progress forwards along the reaction coordinate. The single nucleotide more readily moves in the reverse direction, displaying more freedom of movement or a greater dependence of translocation on self-diffusion. This typically resulted in a longer translocation time for single nucleotides compared to polynucleotides, as illustrated in this example.

The influence of the size of the translocating molecule on progression along the reaction coordinate may also be related to the methodological implementation of ABF (see Figure 2.2): consider the leading atom of the polynucleotide being encouraged to move along the reaction coordinate by the biasing force. As the biased atom moves, the polymeric equilibrium conformation is deformed. The altered conformation of the polynucleotide may begin to relax once the leading atom slows or stops its progression. For example, when the atom moves into a new bin along the reaction coordinate, thus the biasing force is not yet being applied in that new bin. At all times during states of perturbation and relaxation of the translocating molecule, the instantaneous force acting on the biased atom is measured and used to calculate or improve the biasing force. As the biased

atom moves into unexplored bins along the reaction coordinate, the system will evolve from a more equilibrated conformation to a more perturbed one. A freshly perturbed polymeric chain will resist the change by pulling back on the leading atom, giving an increased average instantaneous force in the reverse direction, so the calculated biasing force will be higher in the forward direction. Therefore, as the chain then begins to relax and the pull from the chain reduces, the biasing force will be disproportionately large in the forward direction due to the average instantaneous force for that bin having been largely calculated from the more greatly perturbed state. This effect will encourage consistent progression in the forward direction along the reaction coordinate, since the energetic barriers to translocation are being exceeded by the biasing force, rather than being matched by it.

The phenomenon of translocation progression being influenced by molecular conformation will likely be a factor in all multi-atom molecules, with larger and/or more flexible molecules exhibiting it to a greater degree, resulting in shorter translocation times. It should be noted that this effect results in a deviation from the free energy profile which accurately reflects the free energy landscape; thus it is a systematic error that increases with more slowly relaxing molecules, requiring a larger ζ value to reduce the impact of this effect.

4.2.1.2 Influence on Free Energy Profiles and Force Measurements Per Bin

The impact that the force measurement threshold parameter has on the translocation free energy profiles is explored in this subsection. It is important to determine the non-equilibrium contributions to the free energy estimate, as such contributions can represent deviation from physically meaningful values.

In the absence of the force threshold, the free energy profiles are dominated by non-equilibrium effects resulting from rapid translocation. Figure 4.3 shows the relative heights of single sample free energy profiles from ABF simulations dC_{25} translocation at ζ=0, 5000, 20000 and 80000. The figure shows the considerable difference between the free energy profile when ζ=0 and the profiles at the other three values, exhibiting a 20- to 50-fold increase in the cumulative free energy value after 16 Å of translocation. As discussed in the previous subsection, the difference arises due to the slowly relaxing

conformations of the polynucleotide, which, as this figure shows, have a large impact on the cumulative value of the free energy profile.

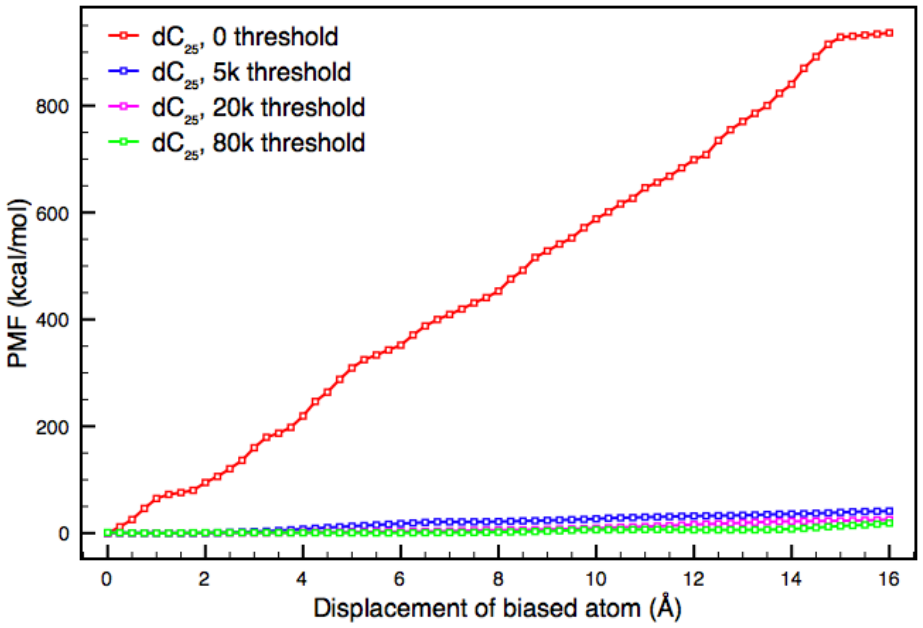


FIGURE 4.3: Free energy profiles from ABF simulations of dC_{25} translocation comparing ζ values of 0, 5000, 20000 and 80000. At higher ζ values, slow relaxing configurations are given more time to relax, resulting in a reduced cumulative free energy value. At a ζ value of 0, the free energy profile is high, whereas the profiles from simulations with 5000, 20000 and 80000 ζ begin to converge within a close region.

Low values of ζ also alter the shape of the free energy profiles. Figure 4.4 shows single sample free energy profiles from ABF simulations of A_{25} and dC_{25} translocation at ζ=0 and ζ=20000. The free energy profiles have been normalised so that the y-axis end points are the same for each profile to illustrate the difference in the profile shapes. The curved shape present when ζ=20000 is entirely absent at ζ=0. This curve is attributed to the influence of the constriction of αHL due to the correlation of the high gradient portions corresponding to molecular configurations of high steric hindrance and electrostatic interactions (demonstrated in the previous chapter); therefore it can be inferred that the ζ=0 profile is dominated by non-equilibrium effects, highlighting the importance of a carefully selected ζ value.

Figure 4.5 shows free energy profiles from single sample ABF simulations of dC_{25} translocation at ζ values of 5000, 20000 and 80000, plotted with histograms of the force measurements per bin used to calculate the biasing force in that bin. The high number of measurements found at the end of some of the force measurement histograms is a result

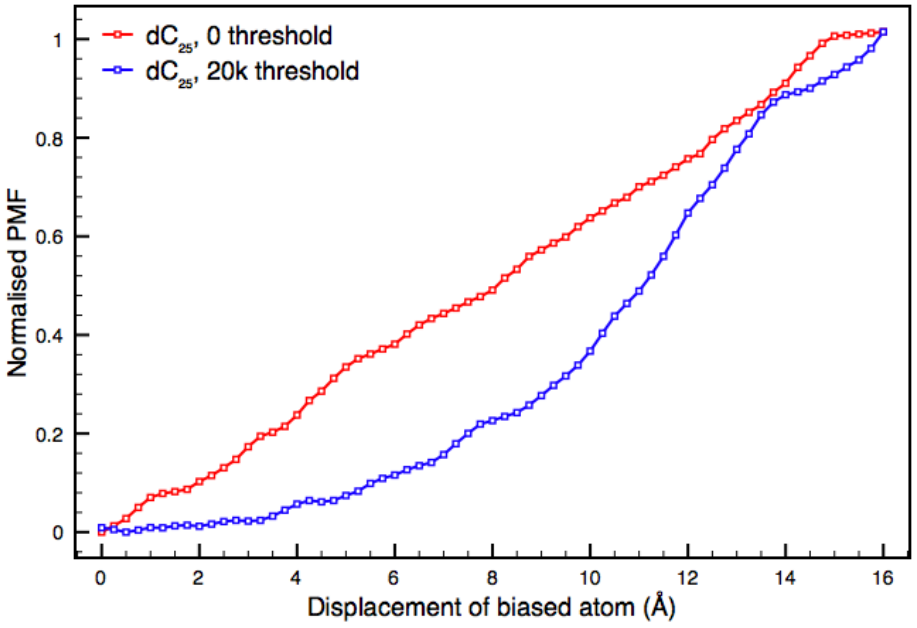


FIGURE 4.4: Free energy plots comparing A_{25} and dC_{25} ABF translocation simulations with ζ values of 0 and 20000. The free energy has been normalised so that the x-axis end points are the same for each profile to illustrate the difference in the profile shapes. The distinctive curved profile shape that is present when ζ is 20000 is absent when ζ is 0.

of the ABF algorithm, which continues to take force measurements after the reaction coordinate has been fully sampled. A fully sampled reaction coordinate will have a free energy estimate for every bin along the length of the reaction coordinate, estimated from a number of instantaneous force measurements greater than the value of ζ. Since the number of timesteps required to fully sample the reaction coordinate is unknown, one must add blocks of timesteps simulated until the reaction coordinate has been sampled; this can lead to additional time spent sampling the end of the reaction coordinate, as observed in these profiles. Variation in the force measurements per bin along the reaction coordinate occur to a degree, the maximum peaks in the histograms tend to be higher for larger ζ values, in a manner which is roughly proportional ζ. Peaks in the force measurements per bin histograms occur with little regularity in terms of peak height and position. While the variation in sampling does not have a dramatic impact on the shape of the free energy profiles, the additional sampling at the end of the reaction coordinate for ζ=5000 and 20000 in Figure 4.5 correlates with a levelling off of the free energy profiles.

As shown in Figures 4.3 and 4.5, the cumulative free energy values begin to converge

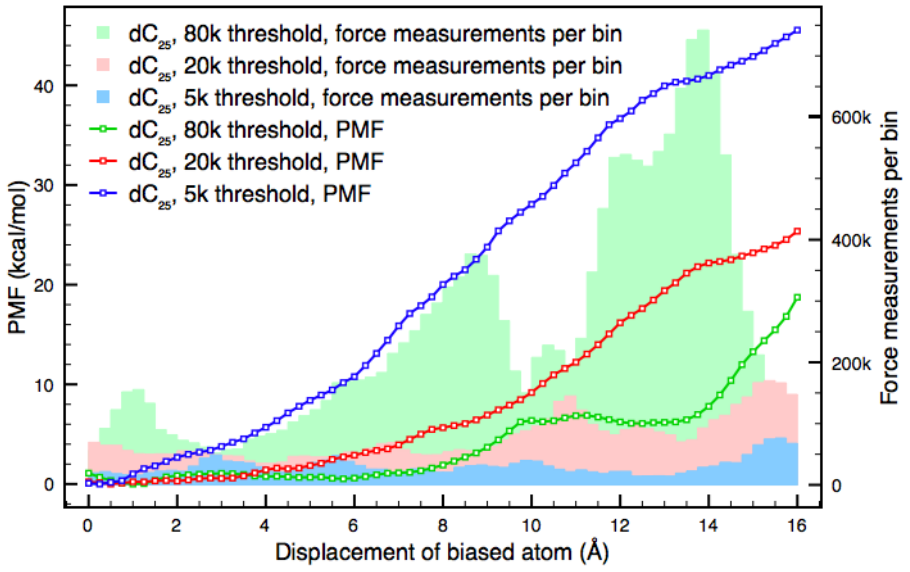


FIGURE 4.5: Free energy profiles from single sample ABF simulations of dC_{25} translocation at ζ values of 5000, 20000, and 80000. The histograms represent the number of instantaneous force measurements per bin. The variation in sampling is roughly proportional to the ζ value.

as ζ increases. The cumulative value where ζ=80000 is assumed to have the lowest systematic error in terms of slowly relaxing conformations. However, at ζ=80000, the shape of the profile exhibits a strong dependance on the sampling per bin. By cross comparison with Figure 4.1, the plateau in the free energy profile between 10 and 14 Å in Figure 4.3 is seen to correspond to a high degree of variance in translocation position in Figure 4.1. It is difficult to determine the precise nature of this effect without more samples to compare; due to the great computational expense of performing simulations at ζ=80000, this has not been performed for this study. For the translocation of polymers using ABF, it may be possible to extrapolate the impact that this systematic error has on the estimated free energy profile by determining the free energy difference across the reaction coordinate at several values of ζ, and extrapolating what the free energy difference might be as $\zeta \to \infty$ (i.e. an unbiased simulation). This may be more difficult with higher ζ values due to the fluctuations in translocation progression and the impact this has on the free energy profile.

Figure 4.6 shows free energy profiles from single sample ABF simulations of dC_1 translocation at ζ values of 5000 and 80000, plotted with the force measurements per bin taken to calculate the biasing force for that bin. Unlike the polynucleotide case, the single

nucleotide free energy profiles show less dependence on the force measurements per bin. The plots display well defined shapes at both values of ζ, regardless of the number of force measurements per bin in the corresponding region. Intuitively we can expect that single nucleotide molecules have faster relaxing degrees of freedom than polynucleotides, accounting for this trend. The plot shows a decrease in the cumulative free energy value as ζ increases, though unlike the polynucleotide case, a significant change in shape is not observed at ζ=80000. A plateau in both of the profiles occurs after roughly 10 Å as was observed in the cv-SMD profiles of the same system. This is in line with the expected influence of the protein constriction on the translocating molecule, as discussed in Section 3.2.3.

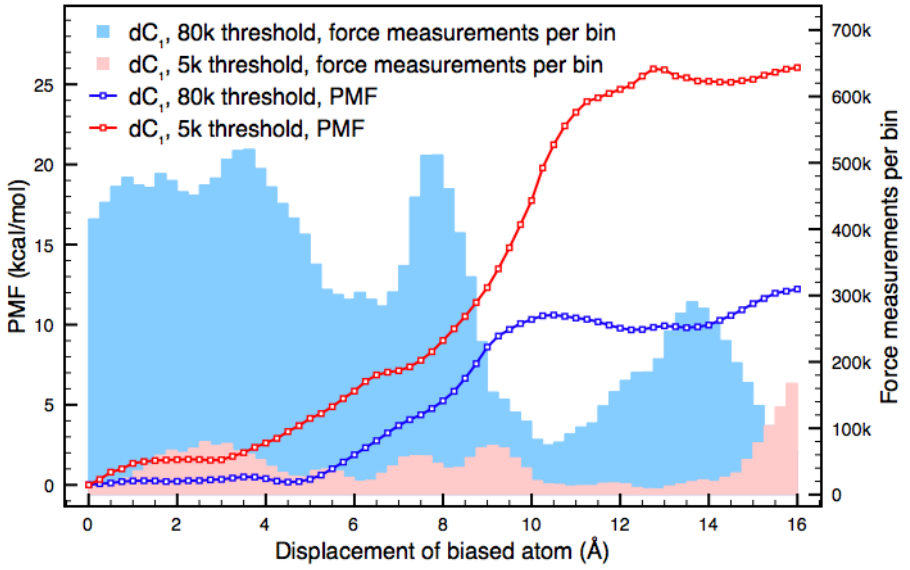


FIGURE 4.6: Free energy profiles from single sample ABF simulations of dC_1 translocation at ζ values of 5000 and 80000. The histograms represent the number of instantaneous force measurements per bin. The shape of the free energy profile is attributed to the influence of the αHL constriction. The shapes of the histograms are not attributed to the influence of the αHL constriction nor the variation in force measurements per bin.

4.2.2 Bin width

Another key parameter is the selection of the bin width, in which instantaneous force measurements are accrued and the resulting biasing force applied. In theory, if the force measurement threshold parameter is scaled with the bin width to ensure that the average sampling per unit distance along the reaction coordinate is kept constant,

then the impact of changing the bin width on the free energy profile should not be overly significant. Figure 4.7 shows free energy profiles and force measurements per bin histograms for dC_{25} ABF simulations with bin widths of 0.1, 0.25 and 1 Å. The ζ value has been scaled accordingly, with ζ values of 2000, 5000, and 20000 respectively. With a bin width of 1 Å and a ζ=20000, the profile is smoother than with smaller bin widths, as one would expect. The force measurements per bin histograms show high variation between bins along the reaction coordinate, the variation for bin widths of 0.25 and 0.1 Å displaying similar relative variation across the reaction coordinate when the plot is normalised so that the y-axis values are scaled to reflect the force measurements per unit distance (not shown). The smallest bin width value of 0.1 Å and ζ=2000 gives a PMF very similar to that at 0.25 Å, though the sampling per bin is the most consistent across the reaction coordinate. The bin widths and ζ values represent a 10-fold change from smallest to largest in these plots, however, the cumulative free energy value varies by $\pm$25%, lying within the expected error bars of each profile. This indicates that changing the size of the bin has a much less significant impact on the PMF than changing the number of force measurements per unit distance along the reaction coordinate.

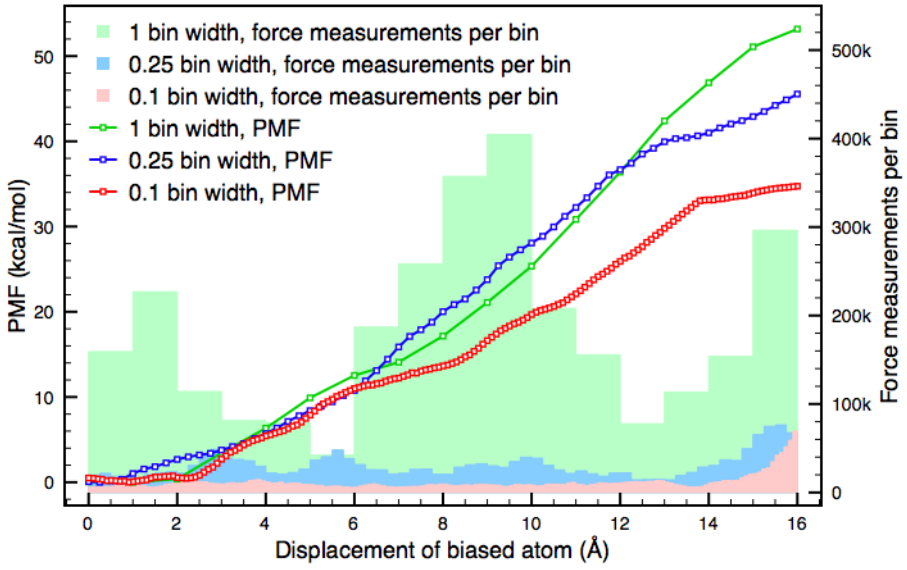


FIGURE 4.7: Free energy profiles from single sample ABF simulations of dC_{25} translocation at bin widths of 0.1, 0.25, and 1 Å. The ζ value has been proportionally adjusted so that a similar number of force measurements per unit distance is performed along the reaction coordinate in each case (2000, 5000, and 20000 respectively). The histograms represent the number of instantaneous force measurements per bin. The bin width has less of an impact on the free energy profiles than the value of ζ.

4.3 Accounting for Errors and Uncertainty

The ABF methodology does not introduce as many sources of error as using cv-SMD in conjunction with JE. Such errors arise from the need to strike a balance between statistical and systematic errors in order to effectively use JE, and from the harmonic spring applied in the cv-SMD methodology (discussed in Chapter 3). ABF circumvents these by directly calculating the forces acting on the biased atom, and applying the biasing force directly into the atom's equations of motion.

As in our data using cv-SMD, the free energy profile error bars were obtained in two forms; firstly, the standard deviation of the accumulated forces per bin (binning error) and, secondly, the sample-to-sample errors when multiple trajectory samples are used in the calculation of the free energy profiles. The calculation of this for the cv-SMD case was discussed in detail in Section 3.1.4. In the ABF case, the binning error calculation was performed by including in the simulation output the system force acting on the biased atom per timestep; these force values were then sorted and averaged, their standard deviation, σ_{bin}, was estimated according to:

$$\sigma_{bin} = \sqrt{\frac{1}{N_d}\sum_{i=1}^{N_d}(x-\bar{x})^2}, \tag{4.1}$$

where N_d is the number of data points, x the value of data point i, and $\bar{x}$ the current mean of x. The binning error was then calculated from σ_{bin} as the standard error:

$$SE_{\bar{x}_{bin}} = \frac{\sigma_{bin}}{\sqrt{N_d}} \tag{4.2}$$

The error from the bin is combined with the error from the preceding bin according to the propagation of errors. This is a necessary step, as each bin free energy value is calculated using the values of the preceding bins. Figure 4.8 shows the binning error plotted as error bars on the free energy profile from a single sample ABF simulations of dC_{25} translocation. As shown, the binning error is negligible in ABF simulations and so will not be considered when examining ABF free energy profiles in the remainder of the chapter. The negligible binning error is due to N_d of Equation 4.2 typically being

of the order of 20,000 to 200,000, since the force measurements are calculated at every timestep.

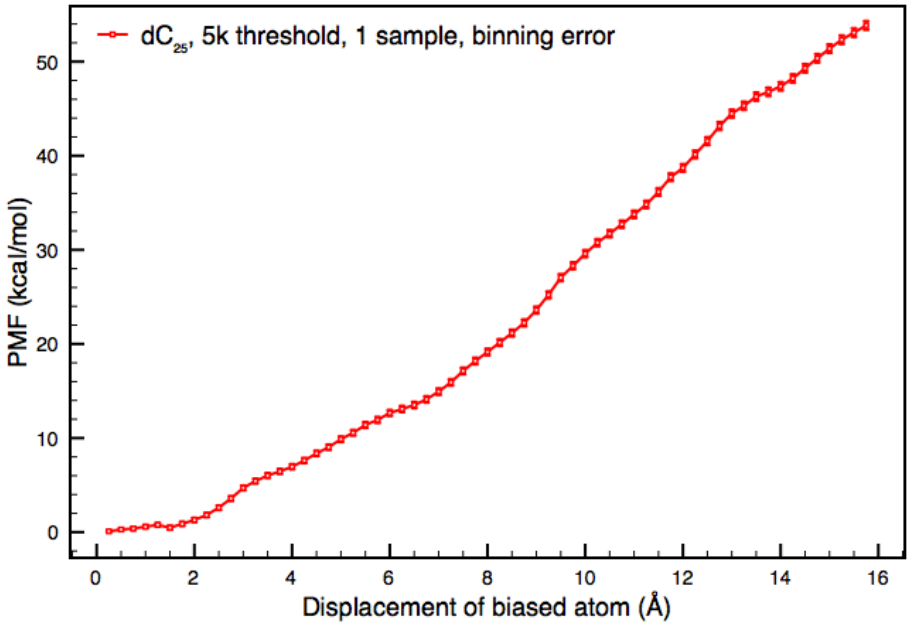


FIGURE 4.8: Binning errors on the free energy profile from a single sample ABF simulation of dC_{25} translocation. The error bars represent the binning errors arising from taking the average of a set of force measurements in each bin, combined with the error from the preceding bin according the propagation of errors. As shown the binning error is very small and so can be negated from error bar calculations.

Secondly, the sample-to-sample variation was estimated using repeat simulations over the full reaction coordinate. Here, four samples of 16 Å translocation were performed, the free energy profile was calculated as the average between them and the variation between samples was estimated according to Equations 4.3 and 4.4. The sample-to-sample error was calculated from the standard deviation:

$$\sigma_s = \sqrt{\frac{1}{N_s - 1} \sum_{i=1}^{N_s} (x - \bar{x})^2} \tag{4.3}$$

Where N_s is the number of samples. The sample-to-sample error was then calculated from σ_{sample} as follows:

$$SE_{\bar{x}_{sample}} = \frac{\sigma_{sample}}{\sqrt{N_s}} \tag{4.4}$$

Examples of sample-to-sample error bars for ABF simulations can be seen in Figure 4.9, 4.10, and 4.11. Here the errors are shown to be significant, resulting in the overlap error bars in some regions of the reaction coordinate. As the binning error is insignificant, the ABF error bars are calculated using Equation 4.4.

4.4 Free Energy Profiles from Adaptive Biasing Force simulations

The trajectory and free energy plots from the previous sections in this chapter were produced from single sample trajectories. These are useful probes but the analysis of such data is limited as single samples may not represent properties such as free energy well (as free energy is calculated from ensemble behaviour). In the following sections, multi-sample free energy profiles are presented, giving an indication of sample-to-sample fluctuation, representing ensemble properties better, and allowing for a comparison with similar cv-SMD simulations that we have performed.

4.4.1 Polynucleotide Translocation

As with the cv-SMD simulations in Chapter 3, the primary goal for these simulations is the reproduction of experimental comparisons of poly(A) and poly(dC) translocation through αHL. This subsection presents multi-sample free energy profiles for the 16 Å reaction coordinates of A_{25} and dC_{25}. As we saw in the cv-SMD investigation, 16 Å is sufficient to examine translocation at the pore-constriction, and doing so allows for multi-sampled data within a reasonable computational budget and time-frame.

It is important to note that the reaction coordinate of the ABF methodology is calculated as a function of distance from the translocating molecule to a reference atom or set of atoms. In cv-SMD, the reaction coordinate is a function of the Cartesian coordinates of the system, if the protein was to move during the simulation, the free energy profile would no longer be an accurate function of the pore interior, therefore, cross-referencing the free energy profile to specific pore residues would not be possible. This necessitates the constraining of the αHL pore in cv-SMD simulations. For ABF simulations, if the reference atoms are reasonably close to the translocating molecule, the free energy

profile will be an accurate function of the length of the pore interior, regardless of the movements of the protein. This aspect of ABF provides an opportunity to leave the protein completely unconstrained, allowing the system to further approach conditions found in experiment.

Leaving the αHL protein unconstrained could have important consequences. In a comprehensive study of the αHL pore in molecular dynamics simulations, the transmembrane barrel was found, on average, to be elliptical in shape [105]. The study found that the pore adopted 7 different orientations of this shape, alternating between them on the scale of several nanoseconds. Constraining the protein prevents this kind of activity, and the consequences this may have on the simulations of this thesis is unknown. Therefore, avoiding the constraining of the protein is an attractive prospect.

A constrained and unconstrained protein provides two different environments, the consequences of which are difficult to predict. Figure 4.9 shows free energy profiles from ABF simulations of A_{25} and dC_{25} translocation through a constrained and unconstrained αHL pore, each with a ζ value of 5000. Each profile is constructed from four samples and the error bars represent the sample-to-sample error. At the end of the 16 Å reaction coordinate in both cases, there is a separation of the A_{25} and dC_{25} profiles beyond the reach of the error bars, with A_{25} showing a higher free energy of translocation. The figure shows the average number of force measurements per bin from the four samples. The profiles show similar qualitative and quantitative end-points whether the protein is constrained or not.

While the simulations of the unconstrained and constrained systems give similar results in terms of their end-points (Figures 4.9(a) and 4.9(b) respectively), they also differ in some respects. Firstly, the histograms for translocation of both polynucleotides are, on average, slightly larger in the unconstrained protein system, showing that the simulations need to be run for more timesteps in order to sample the full reaction coordinate. This is likely a consequence of the protein being allowed to shift, possibly in response to the moving polynucleotide. Such shifts in protein position appear to have a minimal impact on the total free energy difference across the reaction coordinate; as the figure shows, the end points of the profiles in the constrained and unconstrained cases are within error bars of each other. While the profile end-points for each polynucleotide are within error bars for the constrained and unconstrained systems, there is a slight upwards shift for the

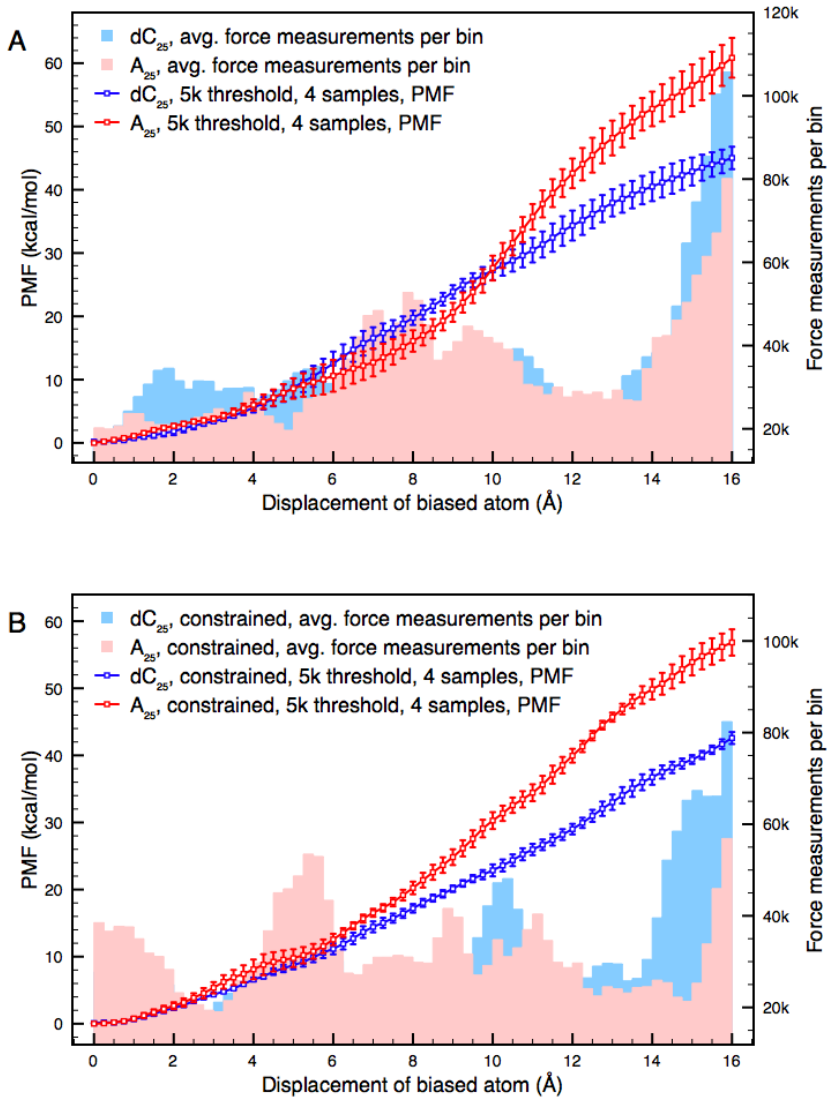


FIGURE 4.9: Global free energy profiles of A_{25} and dC_{25} translocation through wild type αHL from a set of ABF simulations with the protein constrained and unconstrained. The reaction coordinate is from the centre of the alpha carbons of protein residue 111 at the top of the constriction to 16 Å into the transmembrane barrel from that point. The timestep threshold parameter was set to 5,000 for these simulations. Each free energy profile was constructed from four samples, the error bars representing the sample-to-sample variation. The histograms represent the number of instantaneous force measurements per bin. A) Free energy profiles from simulations with the αHL pore completely unconstrained. The free energy profiles are separated beyond the reach of error bars after 11 Å of translocation. At the end of the reaction coordinate, the free energy profile of A_{25} translocation is 46 % higher than that of dC_{25}. B) Free energy profiles from simulations with the C_α atoms of αHL constrained. The free energy profiles are separated beyond the reach of error bars after 6 Å of translocation. At the end of the reaction coordinate, the free energy profile of A_{25} translocation is 33 % higher than that of dC_{25}. Taking an average of the error bars across the whole reaction coordinate, the errors are 69% larger when unconstrained for A_{25}, and 123% larger when unconstrained for dC_{25}.

unconstrained system. Secondly, the error bars are larger for the unconstrained protein. Taking an average of the size of the error bars across the whole reaction coordinate, the errors are 69 % larger unconstrained compared to constrained for A_{25} translocation, and 123 % larger for dC_{25} translocation. When the protein is unconstrained, the scope of the phase space that can be explored becomes greater, therefore the sample-to-sample fluctuations could be expected to be larger for a finite amount of sampling. Thirdly, the separation of the A_{25} and dC_{25} profiles beyond the reach of error bars begins at different points in the constrained system (6 Å) compared to the unconstrained (11 Å). The shapes of the profiles also play a role in this observation; the A_{25} profile shows a curved shape in the unconstrained system, this shape is absent from the constrained system. Given the expected impact of the phosphate-lysine interaction previously demonstrated, this lack of shape could be interpreted as a loss of detail. Finally, the separation of the A_{25} and dC_{25} profiles is larger for the unconstrained system than for the constrained system. At the end of the reaction coordinate, the free energy profile of A_{25} translocation is 46 % higher than that of dC_{25} for the unconstrained system, and 33 % higher in the constrained system.

In deciding whether to constrain the protein or not, one must consider the impact of a shifting protein, both in terms of conformation and overall position. Based on the results in Figure 4.9, the most significant differences in the data give rise to a trade-off between the size of the errors and the accurate recreation of experimental conditions. For the αHL-polynucleotide system at least, the choice does not dramatically impact the end result of the free energy profiles.

As determined from the previous section, simulations where ζ=5000 are expected to contain a significant degree of non-equilibrium contributions in the free energy profiles. At ζ=20000 the non-equilibrium contributions are expected to be much lower. Figure 4.10 shows free energy profiles from ABF simulations of A_{25} and dC_{25} translocation with a ζ value of 20000. The figure shows good separation of A_{25} and dC_{25} polynucleotide free energy profiles, outside the reach of the error bars for the vast majority of the reaction coordinate. The separation between the free energy profiles of the two polynucleotides is the largest per distance travelled of those represented in this thesis. This figure could also be viewed at representing conditions most similar to that found in experiments; given that the average translocation speed is slower as ζ increases, and that the αHL pore in this system is unconstrained. The figure also shows that the error bars are

greater for A_{25} than for dC_{25}, this finding was also observed in cv-SMD simulations in Chapter 3. As shown by the comparison of ζ=20000 and ζ=80000 in Subsection 4.2.1.2, the ζ=20000 profiles here are still likely to represent some non-equilibrium effects. At ζ values higher than 20000, the computational expense in producing multi-sample profiles becomes too great to fully investigate.

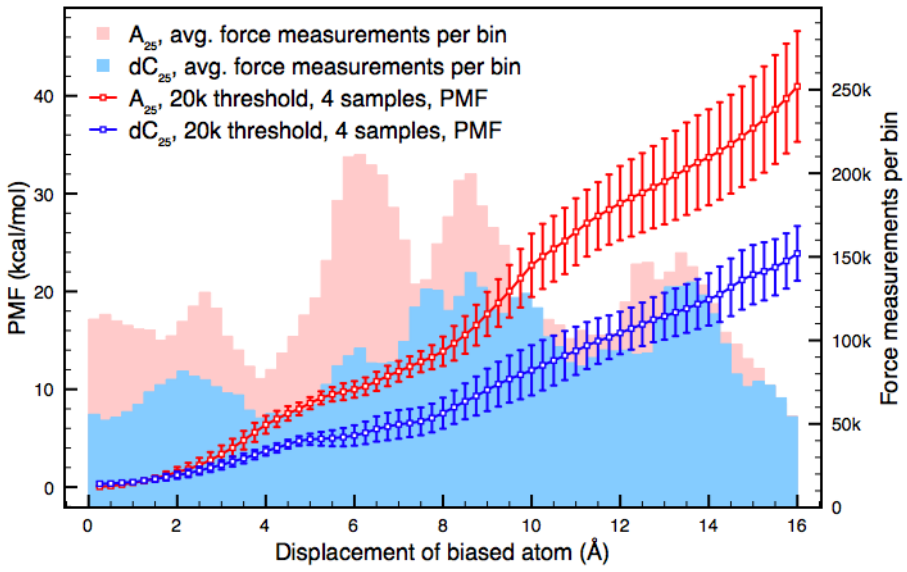


FIGURE 4.10: The global free energy profiles of A_{25} and dC_{25} translocation through wild type αHL from a set of ABF simulations. The reaction coordinate is from the centre of the alpha carbons of protein residue 111 at the top of the constriction to 16 Å into the transmembrane barrel from that point. The timestep threshold parameter was 20,000 for these simulations. Each free energy profile was constructed from four samples, the error bars representing the sample-to-sample variation. The histograms represent the number of instantaneous force measurements per bin. The free energy profiles are separated beyond the reach of error bars after just 3 Å of translocation. At the end of the reaction coordinate, the free energy profile of A_{25} translocation is 71 % higher than that of dC_{25}.

Figure 4.10 also shows the average force measurements per bin from the 4 samples plotted as a histogram. Here, the average force measurements per bin are consistently higher for A_{25} than for dC_{25}. As indicated by the trajectory profiles in Figure 4.2, the size and nature of the translocating molecule is expected to influence the progression along the reaction coordinate and hence influence the force measurements per bin. The fact that a difference in force measurements per bin between polymer types is observed to a greater degree when ζ=20000 compared to when ζ=5000 indicates that progression along the reaction coordinate with the latter parameter value is strongly influenced by

non-equilibrium effects. These non-equilibrium effects appear to dominate the microscopic processes which would otherwise present translocation differences between the two polynucleotides. It is also worth noting that the comparison of the profiles in Figure 4.10 could be considered not completely fair due to the difference in the sampling and the impact that this has on removing non-equilibrium effects. To remedy this, the value of ζ could be increased for dC_{25} or decreased for A_{25}; this would make the average force samples per bin more similar across the reaction coordinate. It is also worth noting that doing so would almost certainly improve the separation between the free energy profiles, as all the data so far in this chapter indicates. It would be interesting to test this hypothesis, but it is unnecessary for the current study as it would provide no new insight while consuming a great deal of computational resources due to the difficulty in predicting values of ζ that would increase or decrease the force measurements per bin by a given amount.

4.4.2 Single Nucleotide Translocation

The comparison of A_1 and dC_1 from Chapter 3 is revisited here, now using the ABF methodology. Given the impact of slow-relaxing forces in the polynucleotide chain, it is important to investigate ABF for smaller molecules such as single nucleotides. Figure 4.11 shows free energy profiles from ABF simulations of A_1 and dC_1 translocation with a ζ value of 5000. The profiles are constructed from four samples per profile and the error bars represent the sample-to-sample error. As indicated by the data in Section 4.2, a higher value of ζ is not as important in reducing non-equilibrium contributions for smaller translocating molecules. The levelling off in the free energy profiles corresponds well to the single nucleotide molecule leaving the confines of the αHL constriction. The strong phosphate-lysine interaction found in cv-SMD simulations for dC_1 is shown to be contributing similarly here as the dC_1 free energy profile shows a higher cumulative free energy than with A_1. The histograms showing the average force measurements per bin are largely similar for both nucleotides.

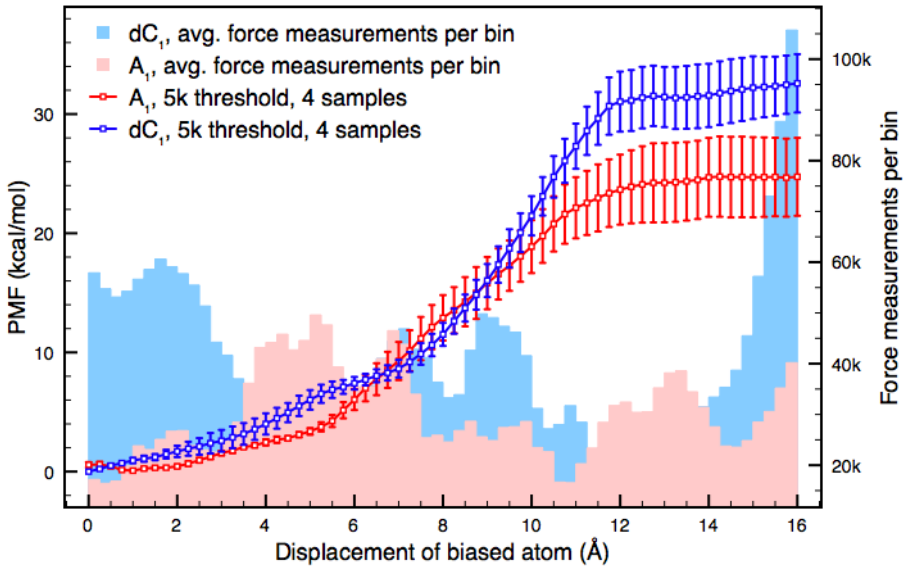


FIGURE 4.11: The global free energy profiles of A_1 and dC_1 translocation through wild type αHL from an ABF simulation. The reaction coordinate is from the centre of the alpha carbons of protein residue 111 at the top of the constriction to 16 Å into the transmembrane barrel from that point. The timestep threshold parameter was 5,000 for these simulations. Each free energy profile was constructed from four samples, the error bars represent the sample to sample variation. The histograms represent the number of instantaneous force measurements per bin. The free energy profiles are separated beyond the reach of error bars after 10.5 Å of translocation. At the end of the reaction coordinate, the free energy profile of A_{25} translocation is 32 % higher than that of dC_{25}.

4.5 Comparison of Constant Velocity-Steered Molecular Dynamics with the Adaptive Biasing Force Method

Chapters 3 and 4 have explored the cv-SMD and ABF methodologies for nucleotide translocation through αHL. Both approaches provided qualitative agreement with the experimental finding that A_{25} experiences greater barriers to translocation than dC_{25}. This section explores which of the two translocation methods is better suited to explore this system, in terms of recreation of experimental conditions and in data quality within a given computational budget. When analysing simulation results in relation to experimental results, quantitative comparisons are difficult to draw without key data such as friction coefficients, and full pore length translocation data. However, qualitative comparisons may be drawn. When considering the simulation pulling methods in relation to each other, we may perform a more rigourous analysis by drawing comparisons between running conditions, error bars, profile shapes, free energy values and required computational resources.

4.5.1 Methodological Comparison

Single channel current recording experiments involve the translocation of a polymer through a protein-pore, this process is effectively a non-equilibrium process, though it is in steady-state due to the constant transmembrane potential. This potential drives the polymer through the pore, and the driving force acts on the entire length of the polymer at all times. The free energy landscape of the solvated and ionised molecular system with respect to the translocating molecule, combined with the applied potential, determines the translocation time (a measurable quantity). So it is this free energy landscape that we wish to estimate using simulations, and the translocation time of poly(A) and poly(dC) is the measurable quantity that we can validate our simulations with, which is itself a reflection of the free energy landscape. Therefore, one key point of comparison between cv-SMD/JE and ABF is how closely the methodology matches the experimental process.

In cv-SMD/JE, the molecule is pulled at non-equilibrium and while the method causes the molecule to be moved at constant velocity, the applied force varies in response to the free energy landscape. The driving force is therefore different to experiment in this way. Another key difference to experiment is that the driving force is applied to the leading atom of the polymer (the reasons for this were discussed in detail in Chapter 3), whereas in experiment it is applied to the whole molecule. During simulations, pulling a polymeric molecule by its leading end can result in deformation from equilibrium conformations [107]. This kind of deformation is expected to occur in experiment due to the dimensions of the pore [49, 107], but as a response to the steric hindrance of the pore, rather than due to being dragged through the solvent. This can be reduced by using a smaller driving force, where relaxation forces have time to act on the molecule. Furthermore, since the reaction coordinate of the ABF methodology is calculated as a function of distance relative to other atoms of reference, the free energy profile will be an accurate function of the length of the pore interior, regardless of the movements of the protein. This allows the protein to be completely unconstrained, permitting the system to further approach conditions found in experiment. In cv-SMD, the reaction coordinate is a function of the Cartesian coordinates of the system. If the protein was to move during the simulation, the free energy profile would no longer be an accurate function of the pore interior. Therefore, cross-referencing the free energy profile to

specific pore residues would not be possible, necessitating the constraining of the αHL pore in cv-SMD simulations.

The two methodologies also differ from each other in several other aspects. Firstly, the ABF reaction coordinate is one-dimensional and therefore it is not restricted in axes orthogonal to the reaction coordinate. cv-SMD, on the other hand, is restricted for those orthogonal axes, and so, assuming a stiff spring constant (needed to use JE), the SMD atom may not stray from a precisely chosen course. In this aspect, ABF is closer to experiment than cv-SMD, where under experimental conditions the molecule is free to explore the full internal dimensions of the pore, and the translocation time is a measure of its transmembrane progression (a one-dimensional quantity). Secondly, the direction of translocation along the reaction coordinate is not consistent in ABF simulations, so deviations from expected conformations can vary from sample-to-sample. Such deviations result in a systematic error in the free energy profiles, as the data from Chapter 3 demonstrated. It is quite straight forward to extrapolate the effect that this has from cv-SMD simulations due to how consistently it will be introduced in proportion to the pulling speed. With data from several pulling speeds, one could extrapolate what the free energy difference would be at infinitesimally small translocation speed, this would be more difficult in ABF. Thirdly, while using cv-SMD/JE requires a balance of statistical and truncation errors in equating the work done to the free energy, as discussed in Chapter 2, ABF involves no such approximations, due to calculating the free energy directly from the system forces, and applying the biasing force directly into the biased atom's equations of motion.

The ABF methodology, while fundamentally different in principle and practice, is also closely related to cv-SMD in certain respects. In ABF for instance, the molecule is permitted to diffuse along the reaction coordinate by a force that is adjusted in response to energetic barriers to translocation. So as in cv-SMD, since the leading molecule is being forcibly relocated, the actual force applied to it scales in response to the energetic barriers to the relocation, the process is not in a steady-state in terms of the driving force. In this way, cv-SMD could be considered more closely related to ABF than it is to constant force-SMD. Additionally, for the polynucleotide-nanopore system, the driving force must also not be applied to the whole polymeric molecule, for the same reasons that cv-SMD may not, as discussed in Chapter 3. This makes the two methodologies

more closely related to each other than they are to methods that use a transmembrane potential as a driving force, for example, Grid-SMD.

4.5.2 Computational Resources and Data Quality Comparison

This subsection compares results from cv-SMD and ABF simulations. In order to draw a fair comparison, the conditions in each methodology have been set to be as similar as possible. The simulation parameters have already been kept largely similar for the simulations so far in this thesis. Such parameters include the Particle Mesh Ewald parameter, the van der Waals energy calculation cut-off, the system temperature etc. Conditions that must be adjusted to produce equality across the methodologies include the average translocation speed, the number of samples used to calculate the profiles, the constraints on the αHL pore, the reaction coordinate bin size, and the total length of the reaction coordinate.

A direct comparison of ABF and cv-SMD for the translocation of A_{25} and dC_{25} is shown in Figure 4.12. Each profile is the average of four sample trajectories, each spanning the full 16 Å reaction coordinate across the pore-constriction. The bin width was set to 0.25 Å and the C_α atoms of the αHL pore were constrained in all instances. The ABF parameters were set to ζ=5000 with a bin width of 0.25 Å, while the cv-SMD parameters were set to a pulling speed of 0.04 Å/ps. These methodology specific parameters equated to roughly the same average translocation speed. The profiles show that both methodologies exhibit a higher free energy profile for A_{25} than for dC_{25}. The cumulative free energy values at the end of the reaction coordinate for ABF are within error bars for cv-SMD for the same polynucleotide.

Figure 4.12 also shows that the two methodologies lead to significant differences in the free energy profiles. The ABF method manifests a greater separation between the free energy profiles of A_{25} and dC_{25} by the end of the reaction coordinate. At the end of the reaction coordinate, the free energy profile of A_{25} translocation is 22 % higher than that of dC_{25} when cv-SMD is used. The difference is 33 % when ABF is used. The separation is also aided by the considerably smaller error bars in the ABF profiles. By contrast, the error bars on the A_{25} and dC_{25} free energy profiles can be seen to overlap in the case of cv-SMD for the majority of the reaction coordinate. Taking an average of the error bars across the whole reaction coordinate, the errors are in the cv-SMD profiles

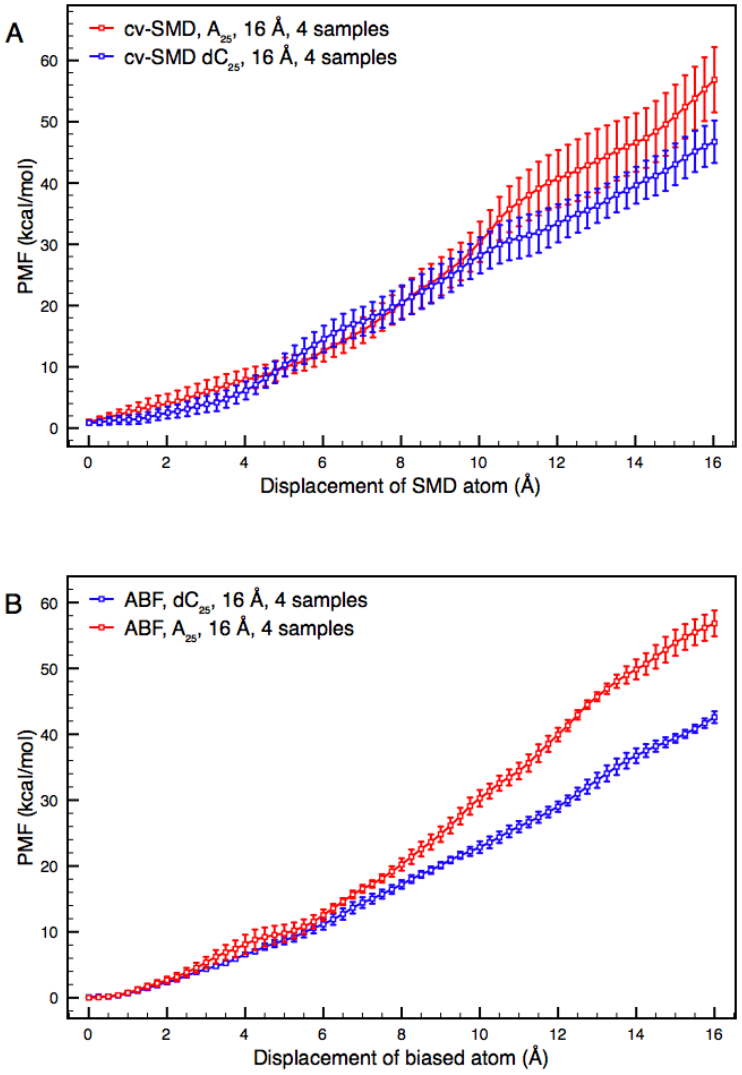


FIGURE 4.12: Free energy profiles from cv-SMD (Figure A) and ABF (Figure B) simulations of A_{25} and dC_{25} translocation under comparable conditions. Each profile is calculated from the average of 4 sample trajectories spanning the full 16 Å reaction coordinate. (A) The ABF parameters were ζ=5000 and the bin width as 0.25 Å. The free energy profiles are separated beyond the reach of error bars after 11 Å of translocation. At the end of the reaction coordinate, the free energy profile of A_{25} translocation is 46 % higher than that of dC_{25}. (B) The cv-SMD parameters were a pulling speed of 0.04 Å/ps and the bin width as 0.25 Å. The free energy profiles are separated beyond the reach of error bars after 15.5 Å of translocation. At the end of the reaction coordinate, the free energy profile of A_{25} translocation is 22 % higher than that of dC_{25}. The two methodologies exhibit a greater free energy profile for A_{25} than for dC_{25}. Compared to the cv-SMD profiles, the ABF data shows greater separation at the end of reaction coordinate, the separation occurs throughout a larger proportion of the reaction coordinate, and the errors are substantially smaller. Taking an average of the error bars across the whole reaction coordinate, the errors are in the cv-SMD profiles are 185% larger than with ABF for A_{25}, and 268% larger for dC_{25}.

are 185% larger than with ABF for A_{25}, and 268% larger for dC_{25}. Data from the full 48 Å reaction coordinate in Chapter 3 shows that in the case of cv-SMD simulations, the free energy profiles begin to separate beyond the extent of error bars after the 16 Å point along the reaction coordinate.

The error bars are likely smaller when using ABF for a number of reasons. Firstly, as demonstrated earlier in the chapter, the binning error is negligible due the large number of measurements taken per bin. Secondly, single samples of ABF, with its translocative motion determined largely by self-diffusion and not by being forced along the reaction coordinate, may be more representative of the true free energy landscape, and so the sample-to-sample fluctuations are lower. Consider that an infinitesimally slowly moving molecule is likely to fully sample all accessible phase space configurations and energy values to a degree which is fully representative of the free energy landscape, and therefore multiple samples of infinitely slowly moving trajectories will have zero sample to sample free energy profile fluctuation. Equally, a fast moving sample will sample less of the accessible phase space, therefore more samples will be required to construct a meaningful free energy profile. It follows then, that a methodology which samples the phase space more effectively will represent the free energy landscape better per sample, and so the sample-to-sample variation will be less. It is feasible that being unconstrained in axes orthogonal to the reaction coordinate also contributes the this effect.

To fully assess the potential of each method, one must also look at the computational cost at comparable conditions, in addition to the quality of the output. In general there is roughly a 3.5% increase in computation time for an ABF simulation compared to a cv-SMD simulation for a fixed number of timesteps with the same number of processors on the same system (tested on the TeraGrid machine Kraken at 576 processors). This is because an ABF simulation must perform additional calculations for the generalised coordinates of the biased and reference atoms, and calculate the average instantaneous force acting on the biased atom. Calculations based on the cv-SMD harmonic spring and the position of the reference atom are comparatively simple, and therefore less computationally demanding.

For the ABF simulations which give rise to the profiles in Figure 4.12, a bin width of 0.25 Å and a ζ value of 5000 is used. This requires roughly 2 million timesteps per sample trajectory at a total cost of roughly 25,000 CPU hours for a four sample profile. Here,

each sample trajectory is produced from two or more simulations in blocks of 100,000 to 1 million timesteps per simulation until the full reaction coordinate is sampled. With a cv-SMD pulling speed of 0.04 Å/ps, for a 16 Å translocation, 2.4 million timesteps are required to perform per sample at a total cost of 29,000 CPU hours for a four sample ensemble average. Here, each sample is produced from four simulations, the combined simulations covering the full reaction coordinate. There is additional simulation time in cv-SMD simulations using the conditions we have used in order to produce the reaction coordinate segment overlap outlined in Section 3.1.2.

It should be noted that a relatively consistent progression along the reaction coordinate for the ABF simulations under these conditions is aided by undesirable slow relaxational non-equilibrium effects. With smaller translocation molecules, or higher ζ values to allow more time for the conformations to relax (thus producing a more correct profile), the number of timesteps required to sample the whole directory would increase and be difficult to predict. As shown in Figure 4.2, where ζ=80000, sampling the reaction coordinate requires roughly 16 million timesteps for a polymeric chain and 20 million timesteps for a single nucleotide. In cv-SMD simulations, the quality of the data may also be improved by slowing down the translocation. In the case of cv-SMD, the increase in computational cost is precise and therefore straightforward to plan and manage.

For the conditions given for this comparison, ABF displays numerous advantages - fewer sources of errors, smaller errors, better separation of free energy profiles, lower computational cost, fewer constraints required, and greater freedom in axes orthogonal to the reaction coordinate (as would be allowed in experiment).

The question remains as to how this comparison would hold up in other systems/conditions. To answer this we must consider individual contributors to each free energy profile. In cv-SMD/JE there are two sources of error from the implementation of the methodology, the harmonic spring and the truncation of the cumulative term in the use of Jarzynski's identity. The latter will likely have a similar contribution in alternative systems, regardless of size or pulling speed. The harmonic spring leads to an increase in the the statistical noise of the output as the harmonic spring constant is increased, yet it must be high enough to approximate a stiff spring. For larger translocating molecules, the spring constant must be scaled up in order to continue approximating a stiff spring, therefore it becomes necessary to introduce more statistical noise. The higher statistical

noise will increase the binning error in the free energy profiles. Therefore, the cv-SMD error would be expected to increase for larger translocating molecules. This scaling of binning error may also be effected by the pulling speed, where faster pulling speeds require higher spring constants in order to approximate a stiff spring, thereby increasing error contribution.

Even if the binning error was completely negated in the cv-SMD profiles, the sample-to-sample contributions to the errors are larger than that of the ABF profiles. This suggests that the ABF methodology samples the free energy landscape more effectively than cv-SMD. This may be surprising as, for the ABF simulations, the reaction coordinate is not restrained in axes orthogonal to it. This lack of restraint increases accessible regions of the phase space, which one would expect to increase the sample-to-sample fluctuations. The opposite case is observed, where each sample appears to represent the free energy landscape well, resulting in low sample-to-sample fluctuations. It is possible that the constrained reaction coordinate in the cv-SMD case imposes certain conformations in the translocating molecule, to a degree which may not be proportionally representative of the ensemble phase space, thereby resulting in more varied individual samples. It is also feasible that the sampling of phase space is improved by the translocative motion in ABF simulations being determined largely by self-diffusion rather than rigidly implemented relocation, again leading to lower sample-to-sample fluctuation. For these advantages in the ABF sampling to be allowed to flourish, the translocating molecule must be allowed sufficient time within each bin along the reaction coordinate, the time spent in each bin would be reduced if the average translocation speed would increase. Therefore at higher speeds, one might expect the sample-to-sample fluctuations to occur to a similar degree in both methodologies, whereas at slower speeds, the ABF methodology would produce better data for a given computational budget. Further investigation at great computational expense would be required in order to fully answer the question of how the ABF and cv-SMD methodologies compare in other systems and/or conditions; it is nonetheless clear that for the translocation of polynucleotides through the αHL protein pore, ABF stands out as the methodology of choice.

4.6 Adaptive Biasing Force Conclusions

Using the adaptive biasing force methodology, free energy profiles were produced from translocation simulations. These simulations covered single nucleotides and polynucleotides translocation through the wild type αHL protein-nanopore.

As with the cv-SMD data, the free energy profiles from ABF simulations of A_{25} and dC_{25} translocation through wild type αHL exhibited the experimentally observed trend that A_{25} experiences greater barriers to translocation than dC_{25}. Unlike with cv-SMD, the ABF methodology allows for the protein-pore to be left unconstrained, permitting conditions more akin to those found in nanopore current recording experiments. By comparing simulations using a constrained and unconstrained αHL pore, it was found that constraining the protein led to smaller errors but smaller separation of A_{25} and dC_{25}. When the simulation conditions were optimised to reduce non-equilibrium contributions, the free energy profiles of A_{25} was shown to be 71 % higher than that of dC_{25} after 16 Å of translocation. Free energy profiles from A_1 and dC_1 translocation through wild type αHL also shared similarities with the cv-SMD results, showing a greater tendency for dC_1 to form a strong phosphate-lysine interaction than for A_1. In the ABF case this has led to the dC_1 free energy profile being higher than that of A_1.

In comparing cv-SMD/JE to ABF for the nanopore-nucleotide problem, results were produced that brought conditions across both methodologies as close as possible. The simulations shared similar molecular models (A_{25} and dC_{25} through a constrained wild type-αHL pore), reaction coordinates, average translocation times, number of samples per profile, and all parameters that were non-specific to the methodology used were kept identical. The resulting free energy profiles from both methodologies were within error bars in terms of their end-point values after 16 Å of translocation. There were also some differences present in the data. The error bars were notably smaller and the separation between the free energy profiles of A_{25} and dC_{25} translocation was larger when using ABF. Given that ABF presents these advantages in data quality, and under our conditions is less computationally intensive for obtaining similar results to cv-SMD, the ABF method is a natural choice for future work of this type.

Additionally, the ABF method has some intrinsic advantages, including the biased atom being permitted more fully explore the internal dimensions of the pore, as one would

expect under experimental nanopore recording conditions. Furthermore, in considering the choice of methodology for other systems, there are many more available forms of reaction coordinate to exploit when using ABF (such as bond angles, RMSD values, and multi-dimensional reaction coordinates). This allows the method to be harnessed for a greater number of problem types. For example, the investigation of free energy barriers to conformational changes relating to a change in bond angle. Additionally, ABF generates fewer sources of error and uncertainty, eliminating the need for cumulant expansions (needed when using JE), and the need for carefully selecting stiff springs (as in cv-SMD). These factors further strengthen ABF as the recommended method. However, cv-SMD a major advantage over ABF in that cv-SMD has a set number of timesteps required to traverse a reaction coordinate distance, therefore allowing precise planning of simulation time and computational budgets.

With ABF established as the preferred method, several nanopore-nucleotide translocation systems present themselves as interesting avenues for future investigation. Firstly, one could change the types of solvation ions used, noting that some nanopore current recording experiments use NaCl, others KCl, and others have mixtures such as NaCl and $MgCl_2$. As we have observed, the cations play a key role in the phosphate-lysine interaction, which is a significant factor in the translocation process. Another problem that could be investigated is the experimental observation that $3'$-led translocation encounters lower energetic barriers than $5'$-led translocation. This system has the advantage of being relatively straight forward to construct. However, the experimental translocation time, and therefore free energy difference, between the differently oriented systems is much smaller than that of poly(A) compared to poly(dC). Therefore, such an investigation would require rigourous sampling for the production of free energy profiles.

Chapter 5

Conclusions

In this thesis, high-performance computer simulations have been applied to investigate the translocation of large biopolymers in atomistic detail. Using constant velocity-steered molecular dynamics and adaptive biasing force simulations, nucleotide molecules were pulled through α-hemolysin protein nanopores, producing free energy profiles of the process. Applying these translocation methods in combination with the use of parallel, high-performance computing systems has permitted translocation distances that cover significant portions of the pore interior, with each reaction coordinate being sampled up to 16 times. The computational strategy produced well sampled free energy profiles and comprehensively analysed errors.

The free energy profiles were in qualitative agreement with experimental nanopore current recordings, showing that polyadenine molecules experience greater energetic barriers to translocation than polydeoxycytosine molecules. The physiochemical nature of the translocation process was investigated using single nucleotide molecules and mutated nanopores. Most notably, it was found that an electrostatic phosphate-lysine interaction at the pore constriction represents a significant energy barrier to nucleotide translocation. By examining the findings from all the aforementioned systems, a major cause for the free energy difference between the two biopolymers was determined to be greater degree of base stacking in polyadenine.

By comparing constant velocity-steered molecular dynamics to the adaptive biasing force method, we have been able to assess the merits of each methodology. While the use of a constant velocity-steering force results in timely, easily budgeted simulations, the

adaptive biasing force method produced better quality data for a lower computational cost, most notably yielding smaller errors. Our findings may be applied to other systems, particularly those with large flexible translocating molecules, when studied in atomistic detail.

One of the potential applications of the nanopores is the sequencing of DNA. Considerable research effort is being directed at devising nanopores that sequence genes quickly and inexpensively to achieve widespread medical benefits. For example, several research groups have demonstrated the potential of synthetic graphene nanopores or α-hemolysin in combination with cyclodextrin and exonuclease for sequencing. The development of new improved nanopores will likely be aided by understanding the fundamental microscopic processes of this system. Using the approach established in this thesis, it is possible to conduct a similar investigation on other nanopores, including synthetic forms, as well as nucleic acids of different composition.

The use of high speed translocation methodologies in combination with high-performance parallel computing has allowed the recreation of a large biologically and experimentally relevant system in atomistic detail. The simulations were run on up to 576 processors per simulation at large computing infrastructure facilities, granting the capability to run multiple samples simultaneously. Parallel-processing has allowed longer reaction coordinates to be explored, while simultaneously performed samples have permitted the production of highly sampled data. With the advent of petascale computing facilities, housing hundreds of thousands of computing cores, it will be possible to produce more rigourously sampled ensembles of data. The consequence of such well sampled data is primarily the increase in reliability of the ensemble averaged values. A higher reliability allows for the investigation of systems that are more difficult to distinguish than those examined in this thesis. As high performance computing technologies continue to develop, it will become possible to perform similar simulations without the need for high speed translocation of nucleic acids, thereby recreating a more accurate representation of the molecular movements and interactions. Furthermore, with greater computing power in place, it will also be possible to apply high speed translocation methodologies to much larger biochemical structures, while retaining the highly valuable atomistic details.

In summary, the work described in the thesis has demonstrated the value in applying high-performance computer simulations to large biochemical structures. The simulations

not only produced useful thermodynamic properties that were linked to experimentally measurable properties, but they also provided unique molecular insight into a system of biophysical and biotechnological interest.

Appendix A

Nanoscale Protein Pores Modified with PAMAM Dendrimers

This chapter reports on a project dedicated to the exploration of covalently bound dendrimers inside αHL nanopores. The results from this project were published in July 2007 in The Journal of the American Chemical Society. The paper, entitled "Nanoscale Protein Pores Modified with PAMAM Dendrimers" [129], is annexed at the end of this chapter. The work for the publication was largely performed in the course of my MSci in Chemistry and during my time as a summer student in 2006 at UCL. For the first three months of my PhD, I performed further experiments on this project. Results produced in that time include the annexed Figure 5, and preparing material for some of the single channel current recording experiments.

The project was performed in co-operation with Helen Kinns, Nick Mitchell, Yann Astier, Rethi Madathil, and Stefan Howorka. The project firstly covered the synthesis of chemically activated PAMAM dendrimers and their characterisation with MALDI-mass spectrometry (MALDI-MS) and high-performance liquid chromatography, and secondly, the generation of PAMAM-modified nanopores and their analysis through sodium dodecyl phosphate polyacrylamide gel electrophoresis (SDS-PAGE) and single channel current recordings. My personal contributions to the complete project were mainly in the the MALDI-MS characterisation of the PAMAM dendrimers and the synthesis and SDS-PAGE analysis of the nanopores.

The aim of our study was to functionally enhance nanopores by covalent modification with PAMAM dendrimers. In general, the use of protein nanopores in single channel current recording experiments can yield information about single molecules. The specificity and fidelity of this information can be improved by engineering the protein nanopore to work in conjunction with a chemically bound molecule or through point mutations. For example, Howorka *et al.* coupled DNA oligonucleotides via an engineered cysteine to αHL. The tethered oligonucleotide was able to specifically sense DNA strands via hybridisation [54]. In another study, Astier *et al.* and Clarke *et al.* mutated cysteine amino acids into the αHL pore to allow covalently couple the molecular adaptor cyclodextrin [34, 55], which allowed for the detection of nucleoside monophosphate molecules. Furthermore, Howorka *et al.* chemically bound poly ethylene glycol (PEG) chains to αHL nanopores [130]. They found that the PEG molecules modulate the ionic current passing through the pore, giving information on the conformational dynamics of the PEG chains. Movileanu *et al.* coupled a biotin molecule at the end of an αHL-bound PEG chain [131]. The biotin was capable of binding to a complementary streptavidin protein, thereby effectively leading to single-molecule detection of protein analytes. Taken together, these studies illustrate that covalent coupling can transfer the functional properties of polymers to nanopores

Our aim was to extend this coupling approach to dendrimers. PAMAM dendrimers are polycationic star-like hyberbranched molecules. As shown in the annexed paper, the coupling of a dendrimer molecule to the inside of a nanopore has two potential uses. The first is that dense dendrimers function as ion-selectivity filters and molecular sieves for the controlled passage of molecules. This is illustrated in Figure A.1. The alteration of pore properties with charged and hyperbranched dendrimers is an approach that might be extended to inorganic nanopores with applications in sensing and separation technology.

The covalent attachment of dendrimers to the inside of a nanopore can also be applied to give information on the channel's internal structure. By using the rigid PAMAM molecule of specific diameter as a probe, one can determine if this molecule is able to bind to a mutated amino acid inside the pore. The accessibility of the mutated amino acid by the probe molecule provides evidence of the internal pore properties or dimensions. Flexible 'soft' polymers such as PEG do not provide the consistent diameter required to distinguish different parts of the same pore. Dendrimers are 'hard'

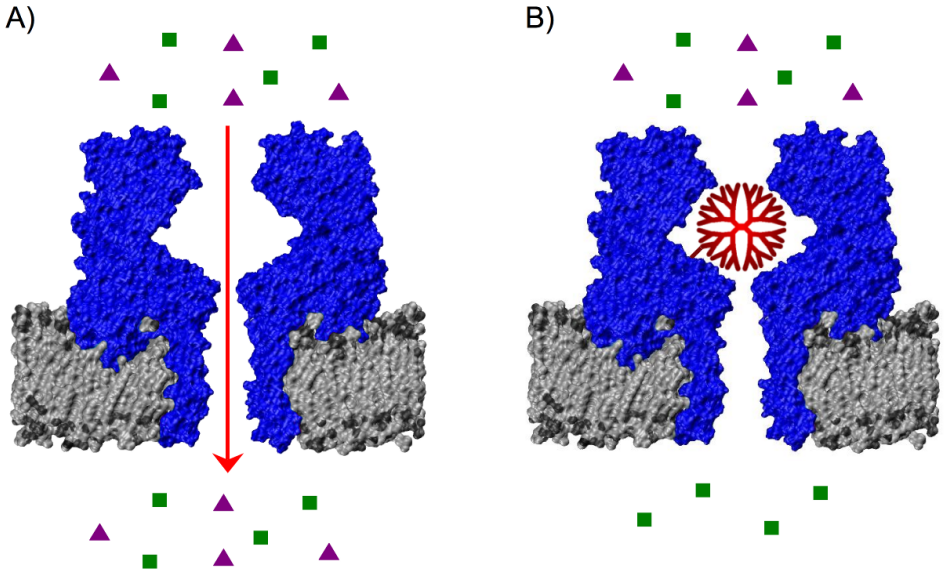


FIGURE A.1: Illustration of αHL modified with a dendrimer molecule, resulting in a molecular sieve. A) An unmodified αHL protein pore (blue) inserted into a lipid membrane (grey) in a single channel current recording experimental setup. Two types of molecules (purple triangles and green squares) lie in solution on the *cis*-side of the membrane. Due to the pore's internal dimensions and properties, both types of molecules are allowed to pass through the pore to the *trans*-side. B) The inside of the αHL protein pore has been coupled to a PAMAM dendrimer (red). The dendrimer can block the passage of larger molecules through the pore due to steric hindrance, or an engineered dendrimer molecule can carry charged terminal groups that can block the passage of molecules with particular electrostatic properties. Here, the passage of the molecules represented by purple triangles has been blocked, resulting in the molecules represented by green squares being selectively allowed through to the *trans-side* of the membrane.

branched polymers that can be engineered at different diameters, proving to be more effective probes than PEG molecules. Our study was able to demonstrate these benefits of PAMAM dendrimers, with the protein pore αHL as an example.

J|A|C|S
ARTICLES
Published on Web 07/18/2007

Nanoscale Protein Pores Modified with PAMAM Dendrimers

Hugh Martin,[†] Helen Kinns,[†] Nick Mitchell,[†] Yann Astier,[‡] Rethi Madathil,[‡] and Stefan Howorka*,[†]

Contribution from the Department of Chemistry, University College London, London WC1H 0AJ, England, United Kingdom and Department of Chemistry, University of Oxford, Oxford, OX1 3TA, England, United Kingdom

Received December 19, 2006; E-mail: s.howorka@ucl.ac.uk

Abstract: We describe nanoscale protein pores modified with a single hyperbranched dendrimer molecule inside the channel lumen. Sulfhydryl-reactive polyamido amine (PAMAM) dendrimers of generations 2, 3 and 5 were synthesized, chemically characterized, and reacted with engineered cysteine residues in the transmembrane pore α-hemolysin. Successful coupling was monitored using an electrophoretic mobility shift assay. The results indicate that G2 and G3 but not G5 dendrimers permeated through the 2.9 nm cis entrance to couple inside the pore. The defined molecular weight cutoff for the passage of hyperbranched PAMAM polymers is in contrast to the less restricted accessibility of flexible linear poly(ethylene glycol) polymers of comparable hydrodynamic volume. Their higher compactness makes sulfhydryl-reactive PAMAM dendrimers promising research reagents to probe the structure of porous membrane proteins with wide internal diameters. The conductance properties of PAMAM-modified proteins pores were characterized with single-channel current recordings. A G3 dendrimer molecule in the channel lumen reduced the ionic current by 45%, indicating that the hyperbranched and positively charged polymer blocked the passage of ions through the pore. In line with expectations, a smaller and less dense G2 dendrimer led to a less pronounced current reduction of 25%. Comparisons to recordings of PEG-modified pores revealed striking dissimilarities, suggesting that differences in the structural dynamics of flexible linear polymers vs compact dendrimers can be observed at the single-molecule level. Current recordings also revealed that dendrimers functioned as ion-selectivity filters and molecular sieves for the controlled passage of molecules. The alteration of pore properties with charged and hyperbranched dendrimers is a new approach and might be extended to inorganic nanopores with applications in sensing and separation technology.

Introduction

Dendrimers are a rapidly expanding class of highly branched globular nanoscale polymers.[1,2] Their synthesis proceeds via either the divergent or the convergent route[3] and offers control over molecular mass, size, shape, degree of branching, and type and number of terminal functional groups. Reflecting their special characteristics, many potential applications have been developed in materials science and nanotechnology for separation technology, surface coatings, and catalysis[4] and in biological sciences for drug and gene delivery, vaccines, and bioimaging.[5] Polyamido-amine (PAMAM) Starburst dendrimers are an important subclass of dendrimers (Figure 1A). They were historically the first dendrimers to be synthesized using the divergent strategy.[1,6] Their synthesis starts with a Michael addition of methyl acrylate to the ethylene diamine core followed by amidation of the tetraester with ethylene diamine, yielding a generation 0 dendrimer. Subsequent Michael addition/amidation cycles provide dendrimers of increasing generation with the number of surface groups doubling each generation. The mass of the polymer also approximately doubles with each extension step until crowding between the terminal surface groups blocks further growth,[7] leading to increasingly rigid compact polymers.[8] The compact structure of PAMAM polymers has been characterized with chemical[9,10] and physicochemical techniques.[11,12]

[†] University College London.
[‡] University of Oxford.

(1) Tomalia, D. A.; Baker, H.; Dewald, J.; Hall, M.; Kallos, G.; Martin, S.; Roeck, J.; Ryder, J.; Smith, P. *Polym. J.* **1985**, *17*, 117−132.
(2) Newkome, G. R.; Yao, Z.; Baker, G. R.; Gupta, V. K. *J. Org. Chem.* **1985**, *50*, 2003−2004. Tomalia, D. A. *Prog. Polym. Sci.* **2005**, *30*, 294−324. Frechet, J. M.; Tomalia, D. A. *Dendrimers and other dendritic polymers*; John Wiley & Sons: New York, 2002.
(3) Newkome, G. R.; Moorfield, C. N.; Vögtle, F. *Dendrimers and Dendrons: Concepts, Syntheses Applications*; Wiley-VCH: Weinheim, 2001.
(4) Frechet, J. M. J. *J. Polym. Sci., Part A: Polym. Chem.* **2003**, *41*, 3713−3725.
(5) Lee, C. C.; MacKay, J. A.; Frechet, J. M.; Szoka, F. C. *Nat. Biotechnol.* **2005**, *23*, 1517−1526.
(6) Matthews, O. A.; Shipway, A. N.; Stoddart, J. F. *Prog. Polym. Sci.* **1998**, *23*, 1−56. Tomalia, D. A.; Frechet, J. M. J. *J. Polym. Sci., Part A* **2002**, *40*, 2719−2728.
(7) Bauer, B. J.; Amis, E. J. In *Dendrimers and other dendritic polymers*; Frechet, J. M., Tomalia, D. A., Eds.; John Wiley & Sons: New York, 2002; pp 255−284.
(8) Li, D.; Tomalia, D. A. In *Dendrimers and other dendritic polymers*; Frechet, J. M., Tomalia, D. A., Eds.; John Wiley & Sons: New York, 2002; pp 285−307. Hierlemann, A.; Campbell, J. K.; Baker, L. A.; Crooks, R. M.; Ricco, A. J. *J. Am. Chem. Soc.* **1998**, *120*, 5323−5324.
(9) Islam, M. T.; Majoros, I. J.; Baker, J. R. *J. Chromatogr. B.* **2005**, *822*, 21−26. Peterson, J.; Allikmaa, V.; Subbi, J.; Pehk, T.; Lopp, M. *Eur. Polym. J.* **2003**, *39*, 33−42.
(10) Subbi, J.; Aguraiuja, R.; Tanner, R.; Allikmaa, V.; Lopp, M. *Eur. Polym. J.* **2005**, *41*, 2552−2558.
(11) Nourse, A.; Millar, D. B.; Minton, A. P. *Biopolymers* **2000**, *53*, 316−328.

10.1021/ja0689029 CCC: $37.00

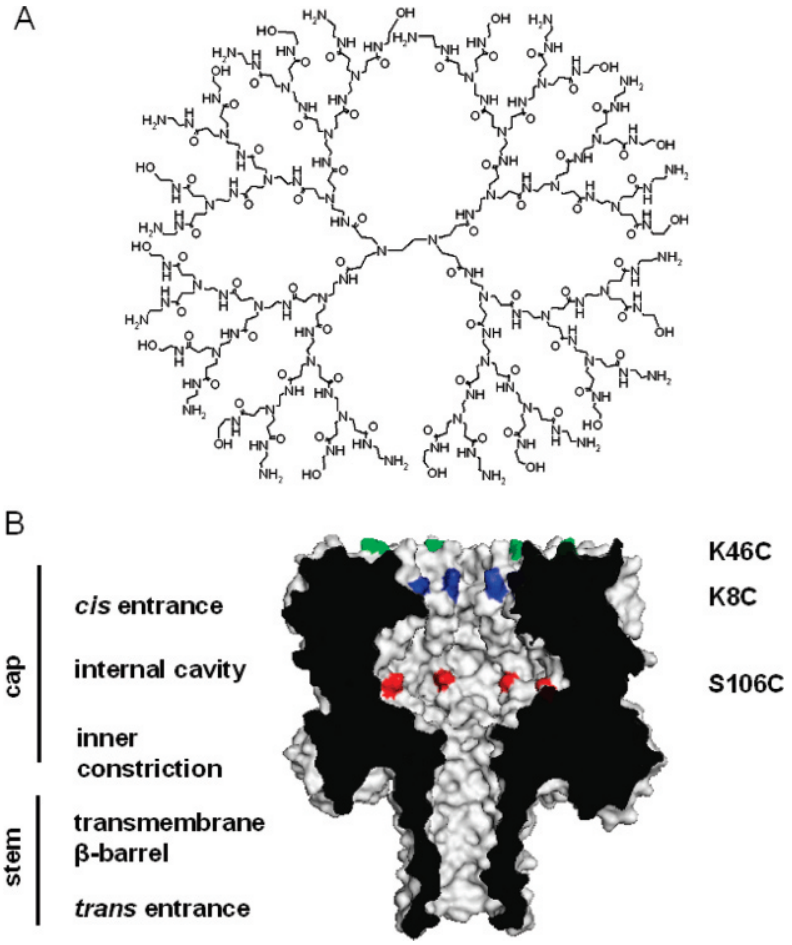


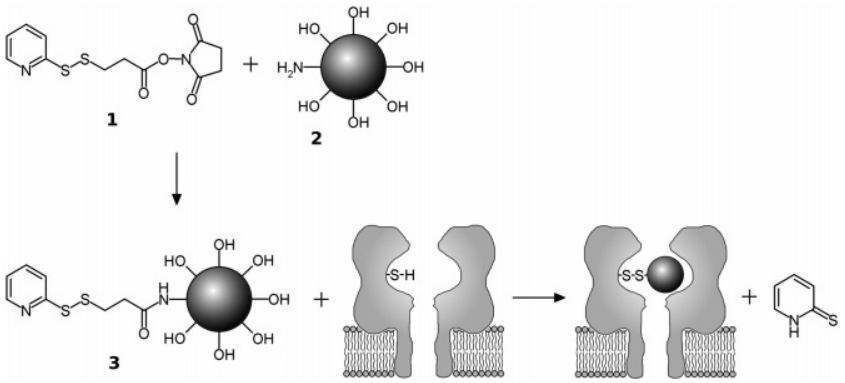
Figure 1. (A) Structure of a polyamido amine dendrimer of generation 3 with 50% amine and 50% hydroxyl terminal groups. (B) Cross-sectional view of the heptameric αHL pore with cysteine substitutions K46C, K8C, and S106C. The model was generated using crystallographic data[13] and PyMol. The internal diameters of the channel are as follows: 2.9 nm, cis entrance; 4.1 nm, internal cavity; 1.3 nm, inner constriction; 2 nm, trans entrance of the β-barrel.

The α-hemolysin (αHL) polypeptide is a bacterial toxin which self-assembles to form a heptameric protein pore. The X-ray structure of the αHL pore resembles a mushroom with a wide cap and a narrow stem which spans the lipid bilayer (Figure 1B).[13] The external dimensions of the heptameric αHL pore are 10 × 10 nm, while the central channel is 2.9 nm in diameter at the cis entrance and widens to 4.1 nm in the internal cavity (Figure 1B). In the transmembrane region, the channel narrows to 1.3 nm at the inner constriction and broadens to 2 nm at the trans entrance of the β-barrel. The defined structure of αHL has facilitated extensive engineering studies and led to the development of tools for the targeted permeabilization of cells[14] as well as new biosensor elements which permit the stochastic sensing of molecules.[15] In stochastic sensing, individual molecules are detected by their ability to modulate ionic current flowing through a single pore. This approach has been used for analytes such as toxic metal ions,[16] drugs,[17] enantiomers,[18] TNT,[19] and nucleotides.[20] Stochastic sensing is also an attractive label-free strategy to study single-molecule kinetics of chemical transformations such as the cis−trans isomerization of azobenzene[21] or the multistep formation or breaking of covalent bonds.[22]

An essential component in these sensors has been the covalent attachment of small molecules and linear polymers within the pore. For example, tethering of DNA oligonucleotides to engineered pores enabled the sequence-specific detection of individual free DNA strands.[23] Organic polymers such as polyethylene glycol (PEG) were also tethered to pores via engineered cysteines. Single-channel current recordings of these pores demonstrated that a single PEG chain modulated the ionic current passing through the pore.[24] On the basis of the

(12) Bosman, A. W.; Janssen, H. M.; Meijer, E. W. *Chem. Rev.* **1999**, *99*, 1665−1688.
(13) Song, L.; Hobaugh, M. R.; Shustak, C.; Cheley, S.; Bayley, H.; Gouaux, J. E. *Science* **1996**, *274*, 1859−1866.
(14) Eroglu, A.; Russo, M. J.; Bieganski, R.; Fowler, A.; Cheley, S.; Bayley, H.; Toner, M. *Nat. Biotechnol.* **2000**, *18*, 163−167.
(15) Bayley, H.; Cremer, P. S. *Nature* **2001**, *413*, 226−230.
(16) Braha, O.; Walker, B.; Cheley, S.; Kasianowicz, J. J.; Song, L.; Gouaux, J. E.; Bayley, H. *Chem. Biol.* **1997**, *4*, 497−505.
(17) Gu, L. Q.; Braha, O.; Conlan, S.; Cheley, S.; Bayley, H. *Nature* **1999**, *398*, 686−690.
(18) Kang, X. F.; Cheley, S.; Guan, X.; Bayley, H. *J. Am. Chem. Soc.* **2006**, *128*, 10684−10685.
(19) Guan, X.; Gu, L. Q.; Cheley, S.; Braha, O.; Bayley, H. *Chem. Biol. Chem.* **2005**, *6*, 1875−1881.
(20) Astier, Y.; Braha, O.; Bayley, H. *J. Am. Chem. Soc.* **2006**, *128*, 1705−1710.
(21) Loudwig, S.; Bayley, H. *J. Am. Chem. Soc.* **2006**, *128*, 12404−12405.
(22) Shin, S. H.; Luchian, T.; Cheley, S.; Braha, O.; Bayley, H. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 3707−3709 and 3523. Luchian, T.; Shin, S. H.; Bayley, H. *Angew. Chem., Int. Ed. Engl.* **2003**, *42*, 3766−3771. Luchian, T.; Shin, S. H.; Bayley, H. *Angew. Chem., Int. Ed. Engl.* **2003**, *42*, 1926−1929.
(23) Howorka, S.; Cheley, S.; Bayley, H. *Nat. Biotechnol.* **2001**, *19*, 636−639. Howorka, S.; Movileanu, L.; Braha, O.; Bayley, H. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 12996−13001. Howorka, S.; Bayley, H. *Biophys. J.* **2002**, *83*, 3202−3210.

Scheme 1. Preparation of Sulfhydryl-Reactive Pyridyldithiopropanoyl PAMAM Dendrimers, 3, from Heterobifunctional Cross-Linker Succinimidyl-3-(2-pyridyldithio)-propanoate (SPDP), 1, and PAMAM, 2, Followed by Coupling To Engineered Cysteines of the Heptameric Protein Pore αHL[a]



[a] For visual clarity, only one of the seven thiol groups of the protein pore is shown. Drawing not to scale.

characteristic current modulations, differences in the conformational dynamics of individual linear polymers of different chain length could be observed.[24] PEG-modified pores also led to the development of biosensor elements capable of detecting protein analytes at the single-molecule level.[25] Given the past work on flexible, "soft" linear polymers, we were interested in placing a "hard" polymer such as a dendrimer into the pore lumen and investigating this approach for its potential in single-molecule studies and development of designed biomaterials.

In this report, we describe the preparation and characterization of sulfhydryl-reactive PAMAM dendrimers and their coupling to cysteine residues of αHL to form an engineered protein pore. PAMAM dendrimers of generations 2, 3, and 5 (G2, G3, and G5) with a mixed surface of terminal hydroxyl/amine groups were modified with the heterobifunctional cross-linker *N*-succinimidyl-3-(2-pyridyldithio)-propanoate) (SPDP) (Scheme 1). The resulting pyridyldithiopropanoyl (PDP) PAMAM reagents were characterized using RP-HPLC, MALDI-MS, SDS gel electrophoresis, and an electrophoretic mobility shift assay based on the formation of a thiol-specific linkage to a cysteine-bearing polypeptide. Dendrimers were coupled to cysteine residues outside and inside the lumen of the heptameric nanoscale αHL pore. The size-dependent coupling of PAMAM into the protein pore was studied because the permeation properties of some dendrimers into narrow nanoscale openings were not known in advance. In this respect, the αHL pore of known X-ray structure served as a reference and calibration standard. The permeation of the compact PAMAM dendrimers was compared to the permeation of a more flexible linear polyethylene glycol derivative. Current recordings of αHL pores carrying a single PAMAM molecule either inside or outside the channel were performed to show that the presence of the dendrimer altered the conductance of the channel in a manner depending on the size and position of the dendrimer. Recordings also demonstrated that PAMAM dendrimers functioned as an ion-selectivity filter and a molecular sieve for the passage of biopolymers. Comparisons of PAMAM protein pores to recordings of PEG-modified pores revealed striking differences, hence demonstrating how the dissimilar structural and dynamic properties of hard and soft polymers affect ion permeation at the single-molecule level.

This study is of interest to the areas of protein research, materials engineering, and single-molecule science. In protein research, sulfhydryl-reactive PAMAM dendrimers can be used as a new type of reagent to study the molecular structure of membrane proteins. Currently used reagents for the substituted-cysteine accessibility method (SCAM)[26] are too small or too flexible[27] for membrane proteins with pore diameters of more than 2 nm. Sulfhydryl-reactive derivatives of rigid PAMAM dendrimers with diameters up to 10 nm overcome these limitations. In materials science, the engineering of the permeation properties of pores or porous structures is an intensive area of research with applications, for example, in warfare sensing and the membrane-based separation of organic molecules and biomolecules.[28] Placing charged and dense dendrimer polymers inside pores represents a new strategy for altering pore properties that can potentially be applied to inorganic structures with the purpose of controlling the passage of molecules based on their charge and/or size. In single-molecule research, current recordings of pores with single dendrimer molecules can reveal dissimilarities to linear polymers, implying that differences in the structural dynamics of compact dendrimers vs flexible linear polymers can be observed at the single-molecule level.

Results

Selection of PAMAM Dendrimers. Dendrimers of different generations were selected to construct αHL protein pores carrying PAMAM outside and inside the channel lumen. The choice of dendrimers was guided by molecular models of αHL and the known dimensions of PAMAM dendrimers. For the preparation of pores with a PAMAM outside the lumen, we selected G5 dendrimer with a hydrodynamic diameter of 6.2 nm and a hard-sphere diameter of 4.2 nm, obtained from the

(24) Howorka, S.; Movileanu, L.; Lu, X.; Magnon, M.; Cheley, S.; Braha, O.; Bayley, H. *J. Am. Chem. Soc.* **2000**, *122*, 2411−2416.
(25) Movileanu, L.; Howorka, S.; Braha, O.; Bayley, H. *Nat. Biotechnol.* **2000**, *18*, 1091−1095.
(26) Karlin, A.; Akabas, M. H. *Methods Enzymol.* **1998**, *293*, 123−145.
(27) Movileanu, L.; Cheley, S.; Howorka, S.; Braha, O.; Bayley, H. *J. Gen. Physiol.* **2001**, *117*, 239−252.
(28) Bayley, H.; Martin, C. R. *Chem. Rev.* **2000**, *100*, 2575−2594. Kohli, P.; Harrell, C. C.; Cao, Z.; Gasparac, R.; Tan, W.; Martin, C. R. *Science* **2004**, *305*, 984−986. Lakshmi, B. B.; Martin, C. R. *Nature* **1997**, *388*, 758−760. Mara, A.; Siwy, Z.; Trautmann, C.; Wan, J.; Kamme, F. *Nano Lett.* **2004**, *4*, 497−501. Li, J. L.; Gershow, M.; Stein, D.; Brandin, E.; Golovchenko, J. A. *Nat. Mater.* **2003**, *2*, 611−615.

solvent exclusion volume.[11,29] The difference between the hydrodynamic and hard-sphere diameter suggests that the dendrimer molecule is less compact than, e.g., a native globular protein of comparable mass.[11,30] On the basis of these considerations, G5 was predicted to stay outside the pore because it is too big to pass the 2.9 nm wide cis entrance of αHL. For the generation of αHL protein pores carrying a dendrimer inside the lumen, we chose PAMAM dendrimers of generations 3 and 2. The hydrodynamic and hard-sphere diameter of G3 is 4.1 and 2.9 nm, respectively,[11,29] while the hydrodynamic diameter of G2 is 2.9 nm (hard-sphere diameter not available). G3 may or may not couple inside the pore lumen depending on whether the permeation across the 2.9 nm wide cis entrance is governed by the hard-sphere diameter (2.9 nm) or the hydrodynamic diameter (4.1 nm). In any case, the smaller G2 with a hydrodynamic diameter of 2.9 nm was predicted to permeate into the pore.

Sulfhydryl-reactive polyamidoamine dendrimers were generated using G5 PAMAM with a mixed surface of terminal −OH and $-NH_2$ groups at an average ratio of 90:10, G3 PAMAM with a mixed surface of 50:50 hydroxyl/amino groups, and G2 PAMAM with terminal NH_2 groups. A mixed surface with hydroxyl groups avoids formation of multimeric aggregates which can be found in purely amino-terminated but not in hydroxyl-functionalized G5 PAMAM dendrimers.[11] The relative percentages of amine groups of the G3 and G5 PAMAM dendrimers were initially chosen to yield approximately the same number of sulfhydryl-reactive groups for each generation of dendrimer (13 for G5 and 16 for G3 and G2).

Synthesis and Chemical Analysis of Sulfhydryl-Reactive Dendrimers. To introduce sulfhydryl-reactive groups, G5, G3, and G2 PAMAM dendrimers were reacted with the heterobifunctional cross-linker SPDP (Scheme 1). The *N*-succinimidyl-activated ester of SPDP couples to the terminal primary amines to yield amide-linked 2-pyridyldithiopropanoyl (PDP) groups. After reaction with SPDP, PAMAM-PDP was extracted with dichloromethane and precipitated in acetone to remove unreacted cross-linker and the hydrolysis product 3-(2-pyridyldithio)-propanoic acid, as confirmed by TLC (see Experimental Procedures, Supporting Information, S4).

RP-HPLC analysis was used to determine the extent to which SPDP had reacted with dendrimers. The samples were compared to unmodified PAMAM that had been subjected to an identical purification procedure. All samples were initially run on a linear gradient from 90:10 H_2O(0.1 wt % TFA)/acetonitrile to 50:50 H_2O(0.1 wt % TFA)/acetonitrile over 30 min. During this gradient, G3-PAMAM was eluted after 9 min compared to the modified G3-PAMAM-PDP that was eluted after 15 min (Figure 2A and B). The increased retention time is in line with addition of hydrophobic PDP groups. The slight broadening of the peaks and appearance of shoulder peaks present for both PAMAM and PAMAM-PDP might reflect structural defects that occurred during synthesis of the dendrimer, for example, by incomplete alkylation of the primary amines or intramolecular cyclization.[10] Unmodified dendrimer starting material was absent from the traces of the pyridyldisulfide-modified dendrimers, indicating

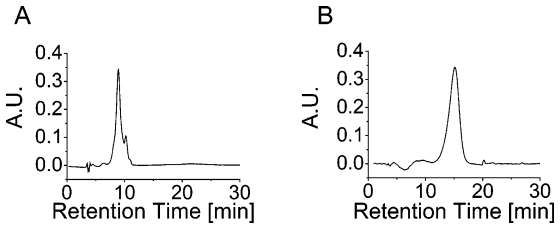


Figure 2. RP-HPLC analysis of the derivatization of PAMAM dendrimer with the heterobifunctional cross-linker SPDP. G3-PAMAM before (A) and after (B) reaction with SPDP.

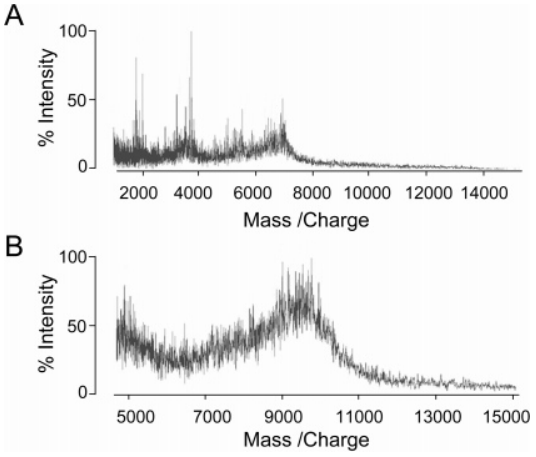


Figure 3. MALDI-TOF analysis of SPDP-modified PAMAM dendrimers to determine the average number of coupled pyridyl disulfide groups. G3-PAMAM before (A) and after (B) reaction with SPDP.

that the reaction had gone to completion. Similar findings on the complete conversion of the starting material were also obtained for G2 and G5 (Supporting Information, S8).

The number of PDP groups coupled to PAMAM dendrimers was determined by MALDI-TOF. Analysis of the unmodified G3 dendrimer gave a broad M^+ peak at 6.9 kD (Figure 3A), which closely corresponds to the theoretical molecular mass of the dendrimer (6909 D).[10] Coupling of the G3 terminal amine groups with SPDP resulted in a shift in the major peak to 9.6 kD (Figure 3B). Each pyridyldithiopropanoyl group has a molecular mass of 199.3 D. Therefore, these data indicate an average of 13 ± 2 pyridyldisulfide groups per dendrimer molecule ($n = 3$; n, number of independent experiments). This is slightly lower than would be expected for complete modification of the 16 amine groups. The same type of MALDI analysis yielded approximately 5.2 ± 1.0 for the G2 and 14 ± 2 PDP groups for the G5 PAMAM dendrimer (Table 1 and Supporting Information, S8 and S9). The number of PDP groups was also determined via photometric analysis, which involved treatment of samples with excess reducing agent dithiothreitol (DTT) to cleave the disulfide bond of PDP, and detection of the cleavage product pyridyl-2-thione at 343 nm. The results of the photometric analysis yielded 5.0 ± 1.5, 11 ± 2, and 13 ± 2 PDP groups for G2-PDP, G3-PDP, and G5-PDP, respectively, which is very similar to the numbers obtained from MALDI-MS (Table 1).

The yield of PAMAM-PDP dendrimers preparations was estimated using sodium dodecylsulfate gel electrophoresis (SDS-PAGE) and Coomassie staining (Figure 4A). The gel band intensities of PDP-modified dendrimers were compared with the intensities of PAMAM dendrimers of a known amount.

(29) Dvornic, P. R.; Uppuluri, S. In *Dendrimers and other dendritic polymers*; Frechet, J. M., Tomalia, D. A., Eds.; John Wiley & Sons: New York, 2002.
(30) Dvornic, P. R.; Tomalia, D. A. In *Polymer Data Handbook*; Oxford University Press: Oxford, 1999; pp 266−270.

Table 1. Chemical Characteristics and Dimensions of G2-PDP, G3-PDP, and G5-PDP

	number of PDP groups in PAMAM-PDP[b]			diameter [nm][c]			
	number of terminal amine groups[a]	MALDI-MS analysis	photometric analysis	hydrodynamic	hard-shell	permeation through 2.9 nm entrance[d]	modification of a highly accessible cysteine residue[d]
G5-PDP	13	14 ± 3	13 ± 2	6.2	4.2	−	+
G3-PDP	16	13 ± 2	11 ± 2	4.2	2.6	+	+
G2-PDP	16	5.2 ± 1.0	5.0 ± 1.5	2.9	na	+	na

[a] Obtained from ^{1}H NMR data. [b] Average of three independent experiments on the photometric detection of pyridyl-2-thione at 343 nm released by treatment of pydridyldithiopropanoyl-PAMAM with the reducing agent DTT. [c] Derived from sedimentation analysis.[11,30] [d] Permeation characteristics through a 2.9 nm pore represent the modification results with αHL heptamer S106C$_7$ and K8C$_7$ (see Supporting Information, S11). For the modification with very accessible residues, heptamer K46C$_7$ was used. Three independent experiments were performed for each modification reaction and gave the same results.

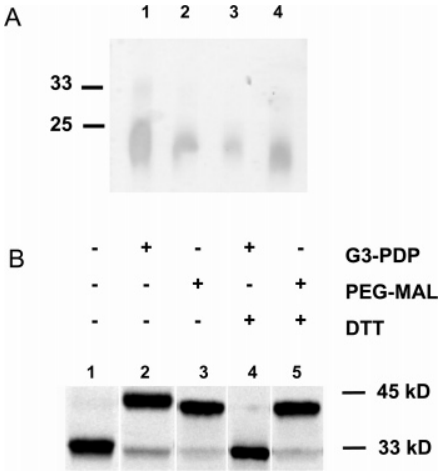


Figure 4. (A) Quantifying the yield of a G3-PAMAM-PDP preparation using SDS gel electrophoresis and Coomassie staining. Solutions of unmodified dendrimers of known concentration and SPDP-modified dendrimers of unknown concentrations were analyzed. The resulting electropherograms were subjected to densitometric analysis to determine the unknown concentrations: lane 1, 67 μM G3-PAMAM; lane 2, 33 μM G3-PAMAM; lane 3, 17 μM G3-PAMAM; lane 4, G3-PAMAM-PDP, concentration determined to be 50 μM. (B) Sulfhydryl-reactive G3-PDP couples specifically to monomeric αHL cysteine mutant K46C and causes a gel shift in SDS-PAGE autoradiographs: lane 1, K46C; lane 2, K46C treated with G3-PDP; lane 3, K46C treated with PEG-MAL 5 kD; lane 4, K46C with G3-PDP and excess reducing agent DTT; lane 5, K46C with PEG-MAL 5 kD and excess DTT.

Densitometric analysis indicated yields of approximately 15% for G2-PDP, 80% for G3-PDP, and 50% for G5-PDP relative to the amount used as starting material (Table 1 & Supporting Information, S10).

PAMAM-PDP Forms a Disulfide Bond with a Cysteine Residue of a Polypeptide Chain. To test their coupling efficiencies, PAMAM-PDP dendrimers were reacted with a water-soluble polypeptide carrying a single engineered cysteine in an accessible position. The [^{35}S]methionine-labeled αHL polypeptide of mutant K46C was generated by coupled in vitro transcription/translation using a cell-free extract. The radiolabeled protein was combined with a 10-fold molar excess of G3-PDP over thiol groups in the reaction mixture and incubated for 20 min at room temperature. The mixture was analyzed by SDS-PAGE and autoradiography. Unmodified αHL polypeptide migrated at a relative molecular mass of 30 kD (Figure 4B, lane 1), while the major protein band after G3-PDP modification was up-shifted to 40 kD (Figure 4B, lane 2). No up-shifted band was observed when the reaction mixture was treated with excess reducing agent DTT (Figure 4B, lane 4), indicating that G3-PDP had coupled specifically to αHL through a disulfide bridge. By quantifying the intensity of the up-shifted αHL-G3 band relative to the unmodified αHL protein band (Figure 4B, lane 2), we deduced that 90% of the cysteine mutant was modified. The same extent of modification was found for G2-PDP and G5-PDP (data not shown). These results confirm that the preparations of SPDP-derivatized PAMAM dendrimers were highly reactive and did not contain contaminating SPDP or the hydrolysis product 3-(2-pyridyldithio)-propanoic acid which might block coupling to cysteines. A similar high extent of modification of 95% was obtained using the sulfhydryl-active reagent mono-methyl polyethylene glycol maleimide 5 kD (PEG-MAL 5 kD) (Figure 4B, lane 3). In line with formation of a thioether bridge between PEG and αHL, treatment with excess DTT did not cleave the conjugate (Figure 4B, lane 5).

A Dendrimer Molecule Can Be Coupled to a Protein Pore Outside the Channel Entrance. Dendrimers G5-PDP and G3-PDP were reacted with the cysteine mutant K46C in an assembled αHL homoheptamer. The seven cysteine residues at the cap of the pore form a ring surrounding the cis entrance (Figure 1B). On the basis of the exposed position of the cysteine residues, both G3-PDP and G5-PDP were expected to react with heptamer. To test the reactivity, radiolabeled homoheptamer K46C$_7$ was generated, treated with G5-PDP, G3-PDP, or, for comparison purposes, PEG-MAL 5 kD, and analyzed via SDS-PAGE and autoradiography to detect the appearance of up-shifted bands. αHL heptamers are not denatured in SDS-PAGE and therefore migrated as defined bands as seen for unmodified K46C$_7$ (Figure 5A and B, lane 1). Upon reaction with G5-PDP, an additional major up-shifted band appeared (Figure 5A, lane 2). This band represents the specific covalent coupling product of G5-PDP and the αHL cysteine mutant as no up-shifted band was observed with unmodified G5 (data not shown). Coupling of G5-PDP to K46C$_7$ produced one but not a second up-shifted band, strongly indicating that coupling of a second G5 dendrimer was disfavored by either steric clash or charge repulsion with the first tethered dendrimer. The first dendrimer could also have coupled to a second or third cysteine residue, thereby preventing reaction with a second molecule. Indeed, some residual cysteine residues of G5-PDP-treated heptamers could still couple to flexible and neutral PEG-MAL 5 kD to produce additional up-shifted bands (Figure 5A, lane 3). A higher extent of modification was obtained when K46C$_7$ was treated with PEG-MAL 5 kD alone (Figure 5A, lane 4). The highly up-shifted and partly unresolved bands represent pore species with 5, 6, or 7 tethered PEG chains as observed in another study.[24] It is worth mentioning that coupling of seven PEG chains with a total mass of up to 7×5 kD $= 35$ kD led to a dramatic gel upshift of the αHL band (Figure 5A, lane 4) while addition of one G5 dendrimer molecule with a comparable mass of 29 kD produced a lesser gel shift (Figure 5A, lane 2) even though addition of

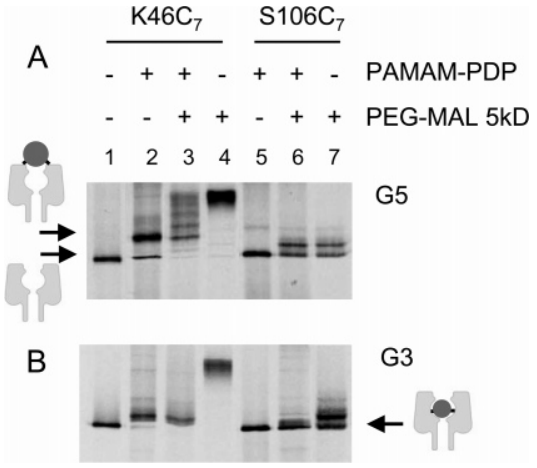


Figure 5. Coupling of a PAMAM-PDP dendrimer inside and outside the lumen of a protein channel. Homoheptameric αHL cysteine mutants were reacted with G5-PDP, G3-PDP, and PEG-MAL and analyzed via gel electrophoresis and autoradiography. $K46C_7$ (lanes 1−4) and $S106C_7$ (lanes 5−7) were treated with PAMAM-PDP (lanes 2 and 5), PAMAM-PDP followed by PEG-MAL (lanes 3 and 6), or PEG-MAL (lanes 4 and 7). The modification reactions were performed with G5-PDP (A) or G3-PDP (B).

the positively charged and unfolded PAMAM polymer would have been expected to slow down the electrophoretic migration of the protein. This difference between PEG and PAMAM likely reflects the different ways in which flexible vs compact polymers interact with the polyacrylamide gel meshwork. Most probably, PEG chains became entangled within the holes of the meshwork, resulting in a retarded migration of the protein band, while PAMAM is too compact to get intertwined.

Reaction of $K46C_7$ heptamers with the smaller G3-PDP resulted in two up-shifted and closely migrating bands (Figure 5B, lane 2 and Supporting Information, S11 for a magnified image of the double band), which likely represent heptamers with one and two G3 dendrimers. In line with the lower molecular mass of G3 (M_r 6.9 kD), both G3-heptamer conjugates migrated lower than the conjugate with G5-PDP (M_r 28.9 kD) (Figure 5A, lane 2). Interestingly, reaction of G3-modified $K46C_7$ heptamers with PEG-MAL 5 kD did not produce additional up-shifted bands (Figure 5B, lane 3), which were however observed in the case of G5-modified heptamers (Figure 5A, lane 3). This indicates that two G3-PDP dendrimers reacted with most of the cysteines, thus blocking further coupling to the flexible PEG-MAL polymer.

A Single PAMAM Dendrimer Can Be Attached Inside the Lumen of a Protein Pore. After establishing that G5-PDP and G3-PDP coupled to very accessible residue K46C, we examined position S106C which is located inside the internal cavity of the αHL pore (Figure 1B). First, we tested the reactivity of the engineered cysteine residue with PEG-MAL 5 kD. Gel electrophoretic analysis of radiolabeled $S106C_7$ heptamers revealed one up-shifted band (Figure 5A, lane 7), confirming successful modification with the polymer. A single and not multiple up-shifted band was obtained because the narrow pore accommodates only one PEG 5kD molecule as found in theoretical[31] and other experimental studies.[24,25] Next, we tested the reaction of $S106C_7$ with the dendrimeric reagents. On the basis of the known hydrodynamic and hard-sphere diameters of the PAMAM dendrimers, we did not expect G5-PDP to permeate into the lumen. By contrast, G3-PDP may couple to the cysteine residues in the internal cavity provided its permeation across the 2.9 nm wide cis entrance is governed by the hydrodynamic diameter (2.9 nm) rather than the hard-sphere diameter (4.1 nm). Gel electrophoretic analysis revealed that reaction with neither G3-PDP nor G5-PDP led to a major up-shifted band (Figure 5A and 5B, lane 5). The lack of up-shifted bands upon PAMAM-PDP addition suggests that $S106C_7$ did not react with the dendrimers. Another possible explanation is that coupling occurred but failed to produce an appreciable gel electrophoretic shift as the tethered dendrimer would reside inside the lumen of the pore. To discriminate between these two possibilities, we treated the $S106C_7$ heptamer with G5-PDP followed by PEG-MAL. Reaction with the flexible PEG chain is known to yield an up-shifted band and can therefore be used to probe whether S106C is accessible or the internal cavity is blocked by a dendrimer molecule. Gel electrophoretic analysis revealed that dual treatment by G5-PDP and then PEG-MAL 5 kD led to an up-shifted band (Figure 5A, lane 6), implying that G5-PDP did not permeate into the internal cavity. By contrast, reaction with G3-PDP and then PEG-MAL 5 kD did not yield an up-shifted band (Figure 5B, lane 6), indicating that G3 permeated into the pore and coupled to S106C, thereby blocking subsequent modification with PEG-MAL. A similar finding on the differential accessibility of G5-PDP, G3-PDP, and PEG-MAL 5 kD was obtained by testing the reactivity of $K8C_7$, which is positioned at the 2.9 nm wide cis entrance (Figure 1B and Supporting Information, S11). These results demonstrate that G3 but not G5 permeated through the cis entrance into the internal cavity. The finding that G5 did not enter the pore is remarkable considering that PEG-MAL 10 kD whose hydrodynamic diameter of 6.2 nm is identical to G5-PAMAM permeated into the pore to couple to residue S106C (Supporting Information, S12). This highlights the different degrees of structural flexibility of dendrimers vs flexible PEG polymers.

Single Dendrimer Alters the Current Flow Through a Protein Pore. The presence of a PAMAM molecule inside the lumen of the heptamer was expected to alter the passage of ions through the channel. We used single-channel current recordings to investigate to what extent the ion conductance was influenced. The recordings were performed with heptamers $S106C_7$, $S106C_7$-G3-PAMAM, $S106C_7$-G2-PAMAM, and $K46C_7$-G5-PAMAM, which had been eluted from preparative PAGE gels. At a potential of +100 mV (chamber on the cis side of the protein grounded), current traces of unmodified $S106C_7$ exhibited a unitary conductance of 925 ± 73 pS ($n = 7$) and an rms noise of 1.82 ± 0.17 normalized to the noise at 0 mV (Figure 6A). By contrast, traces of $S106C_7$ with a single G3-PAMAM inside the lumen had a lower current (Figure 6B) with an average conductance of 512 ± 63 pS and a normalized rms noise of 2.82 ± 0.33 ($n = 8$). The current blockade of 45 ± 6% relative to the open channel indicates that the presence of the dense dendrimer inside the pore lumen largely blocked the permeation of ions by either steric or electrostatic effects or a combination of both. Placing a smaller and less dense G2 dendrimer led to a less pronounced current reduction of 25 ± 5% (Figure 6C) (conductance of 686 ± 114 pS, normalized rms noise of 4.35 ± 0.48, $n = 5$).

The altered channel properties are specific for the presence of dendrimer because addition of excess reducing agent DTT

(31) Kong, C. Y.; Muthukumar, M. *J. Am. Chem. Soc.* **2005**, *127*, 18252−18261.

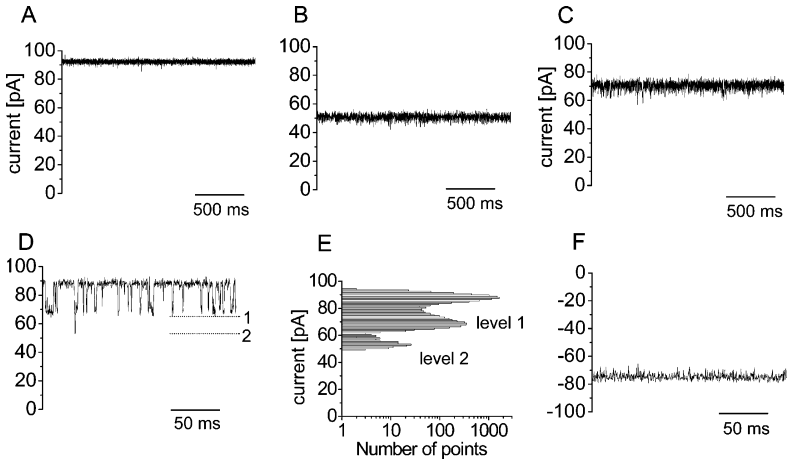


Figure 6. Representative single-channel current traces of (A) $S106C_7$, (B) $S106C_7$-G3-PAMAM, (C) $S106C_7$-G2-PAMAM, and (D and F) $K46C_7$-G5-PAMAM. (E) All-points histogram of $K46C_7$-G5-PAMAM (D) with a duration of 2 s. The recordings were performed in 1 M KCl, 20 mM Tris·HCl pH 7.5 at a transmembrane potential of +100 mV (A−D) or −100 mV (F) with the chamber on the cis side of the protein pore grounded. The currents were filtered at 1 kHz and sampled at 5 kHz.

Table 2. Permeability Ratios for αHL Heptamers Modified with PAMAM Dendrimers

				properties of PAMAM inside protein pore	
heptamer	V_r [mV][a]	P_{Cl^-}/P_{K^+}[a]	P_{K^+}/P_{Cl^-}[a]	hydrodynamic diameter	no. of free amino groups on surface[b]
$S106C_7$	−4.5[c]	1.47 ± 0.06	0.68 ± 0.03	na	na
	5.4[d]	1.59 ± 0.06	0.62 ± 0.2	na	na
$S106C_7$-G2	−9.9[c]	2.41 ± 0.14	0.41 ± 0.02	2.9	11
	11[d]	2.69 ± 0.25	0.37 ± 0.03		
$S106C_7$-G3	−4.9[c]	1.52 ± 0.06	0.65 ± 0.03	4.2	3
	6.6[d]	1.77 ± 0.05	0.56 ± 0.02		

[a] Average of at least three independent experiments ± standard deviation. [b] Determined by subtracting the number of amide-coupled PDP groups from the total number of primary amino groups prior to modification with SPDP. [c] 300 mM KCl (cis), 100 mM (trans). [d] 100 mM KCl (cis), 300 mM (trans).

cleaved the disulfide bonds between the cysteine and PAMAM upon which the conductance state rose to the open state value (data not shown). After examining the conductance changes caused by dendrimers inside the pore lumen, we investigated the effect of G5-PAMAM tethered to K46C just outside the lumen. A typical trace of $K46C_7$-G5-PAMAM is displayed in Figure 6D. The trace with a conductance of 880 pS was decorated with fast downward current fluctuations. An all-points-current histogram generated from a trace of $K46C_7$-G5-PAMAM with a duration of 2 s displays the current levels for the fluctuations (Figure 6E). The peak at 88 pA corresponds to the open channel, while the peak at 68 pA and a minor peak at 53 pA reflect blockade level 1 and blockade level 2, respectively. Blockades to level 2 were not investigated further due to their low frequency. The dominating current blockades from the open channel to level 1 were characterized by an amplitude of 210 ± 23 pS and an average duration of 2.2 ± 0.4 ms. The events occurred at a high frequency of 150 ± 30 s^{-1} ($n = 3$, total of 2000 events). The current fluctuations of $K46C_7$-G5-PAMAM only occurred at a potential of +100 mV but not at −100 mV (Figure 6F), suggesting that the events were dependent on movement of the charged PAMAM relative to the polarity of the potential.

PAMAM Inside the Pore Lumen Acts as an Ion Selectivity Filter and Molecular Sieve. To test the utility of coupling a dendrimer inside a pore, we examined how PAMAM altered the movement of ions through the modified αHL pores. The permeability ratio P_{Cl^-}/P_{K^+} was determined for heptamers $S106C_7$, $S106C_7$-G2-PAMAM, and $S106C_7$-G3-PAMAM. $I-V$ curves were constructed for currents recorded with both cis/trans and trans/cis KCl gradients (Supporting Information, S13), and charge selectivities were calculated from the reversal potential, V_r (Table 2). For $S106C_7$, this analysis yielded a permeability ratio of $P_{Cl^-}/P_{K^+} = 1.47 \pm 0.06$ ($n = 3$), implying that αHL is a weakly anion-selective channel in agreement with previous studies.[32] For $S106C_7$-G2-PAMAM, the permeability ratio was found to be 2.41 ± 0.14 (Table 2), suggesting that PAMAM made the pore more anion selective. This preference for anions is most likely due to the 11 additional protonated positively ionized primary amino groups on the surface of the dendrimer inside the channel lumen. By comparison, $S106C_7$ pores modified with G3-PAMAM displaying only three terminal primary amino groups had a P_{Cl^-}/P_{K^+} ratio of 1.52 ± 0.06 (Table 2), which is very similar to the value for the unmodified pore. The G3 dendrimer had on average only three primary amino groups because the other amines were converted into amide bonds in the course of the attachment to the PDP moiety.

The permeability ratio for $S106C_7$-G2-PAMAM of 2.41 might be further enhanced by increasing the number of positive charges in the dendrimer. It can be expected that PAMAM dendrimers with more surface primary amines will lead to a bigger preference for anions. Quaternary amines[33] could also be used to generate permanent positive charges. In addition, the preferred passage of anions over cations might also be enhanced by changing the size of the dendrimer. Currently, pore-lodged G2 PAMAM with a diameter of 2.9 nm is 1.2 nm smaller than the

(32) Gu, L. Q.; Dalla Serra, M.; Vincent, J. B.; Vigh, G.; Cheley, S.; Braha, O.; Bayley, H. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 3959−3964.
(33) Lee, J. H.; Lim, Y. B.; Choi, J. S.; Lee, Y.; Kim, T. I.; Kim, H. J.; Yoon, J. K.; Kim, K.; Park, J. S. *Bioconjugate Chem.* **2003**, *14*, 1214−1221.

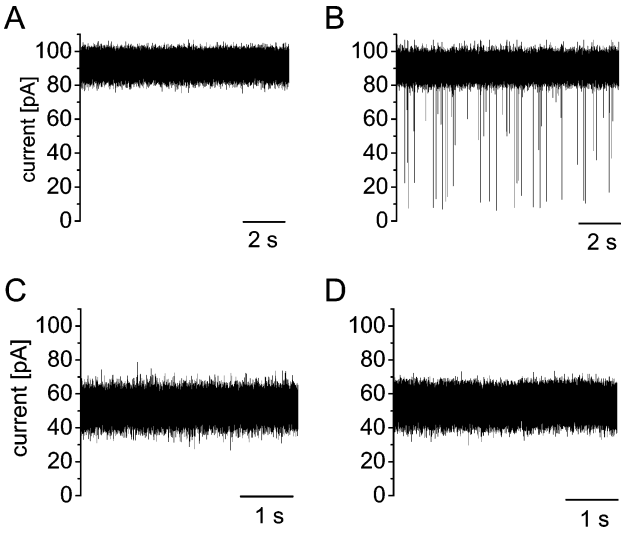


Figure 7. PAMAM acts as a molecular sieve for the passage of RNA polymers through the αHL pore. Single-channel current traces of (A) $S106C_7$, (B) $S106C_7$ with 1 μM RNA oligonucleotide C_{30} at the cis side, (C) $S106C_7$-G3-PAMAM, and (D) $S106C_7$-G3-PAMAM with 1 μM RNA oligonucleotide C_{30} at the cis side with a potential of +100 mV. The traces were filtered at 10 kHz and sampled at 50 kHz.

4.1 nm wide internal cavity. The resulting gap between the dendrimer surface and the channel wall is probably big enough to accommodate the major flow of ions around rather than through the dendrimer. Minimizing the gap is expected to force ions to pass closer to or through the dendrimer and hence increase the effect of positive dendrimer charges on ion flow.

After establishing that PAMAM dendrimers can alter the flow of small ions through the pore, we examined to what extent the dendrimer would affect the permeation of larger molecules. In particular, we tested whether PAMAM would function as a molecular sieve for translocation of oligoribonucleotides. The experiments were conducted with heptamers $S106C_7$ and $S106_7$-G3-PAMAM and oligoribonucleotide C_{30}. When 1 μM C_{30} was added to the cis side of $S106C_7$ with a potential of +100 mV at the trans side, frequent short current deflections appeared (Figure 7B) which were, however, absent in recordings without RNA (Figure 7A). The high-amplitude blockades had an average current amplitude of 87 ± 4% of the open-channel conductance equivalent to a residual conductance level of 120 ± 30 pS and a dwell time of 133 ± 34 μs and occurred at a frequency of 6.5 ± 1.2 s^{-1} ($n = 3$), which are similar to the findings from other nanopore recordings of oligoC-RNA.[34] The high-amplitude events represent the translocation of nucleic acids strands from the cis to the trans side of the pore which temporarily blocks movement of ions through the narrow inner constriction.[34] By contrast, current traces of heptamer $S106C_7$-G3-PAMAM before (Figure 7C) and after addition (Figure 7D) of 1 μM C_{30} did not show any apparent differences. Analysis of extended C_{30} traces revealed that the frequency of occurrence of high-amplitude events for C_{30} traces was 0.005 s^{-1}, which is close to the background noise of the recordings without RNA. The absence of high-amplitude events suggests that the dendrimer blocked the passage of the nucleic acids strands. It is very unlikely that translocation of RNA molecules occurred while their current signatures were masked by the lower conductance level of $S106C_7$-G3-PAMAM pores. If they had occurred, RNA translocation events with a residual conductance of 120 pS should have given rise to highly distinctive current deflections clearly visible in traces of $S106C_7$-G3-PAMAM with a mean conductance of 512 ± 63 pS (Figure 7D). Our data illustrate that a PAMAM dendrimer acts as a size-dependent molecular sieve for the passage of matter through the pore. G3 PAMAM reduced the flux of potassium (M_w 39 D) and chloride ions (M_w 35.5 D) by 45% compared to unmodified pores, while the translocation of RNA polymers (M_w 8614 D) was reduced by more than 3 orders of magnitude. We note that the observed molecular weight cut off is specific for the PAMAM dendrimer of generation 3 and a potential of +100 mV.

Interpretation of Conductance Properties of PAMAM-Modified Pores and Comparison to Pores Modified with PEG. Current traces of dendrimer-modified pores provided insight into the biophysical nature of engineered constructs. For example, recordings from αHL pores with internal G2 and G3 dendrimers established a direct and positive dependence between the size of the polymer inside the pore lumen and the extent of the current blockade. In addition, comparison of G2 and G3 traces with those of G5 revealed information about the conformational freedom of the dendrimer located inside and outside the narrow channel lumen, respectively. In these pores there were no large current steps, indicating that large conformational rearrangements of the dendrimer did not occur most likely due to the steric restraints in the narrow lumen. However, the normalized rms noise of G2 (4.35 ± 0.48) and G3 (2.82 ± 0.33) traces was higher than that of unmodified pores (1.82 ± 0.17), suggesting some minor movement of the dendrimer relative to the pore and/or small conformational changes within the dendrimer structure. Placing G5 outside the pore removed some of the steric constraints, and in line with our simple model, larger current fluctuations occurred in the traces indicating a movement of the dendrimer. The fluctuations oscillated at a rate of 150/s between the conductance level for the open channel and a 22% blockade level. These fluctuations can be interpreted as a rattling movement of the dendrimer between a state very close to the pore entrance and a state further away. The molecular forces for the fast dynamic changes could be the voltage-driven movement of the positively charged dendrimer outside the pore (positive potential at trans side) and the entropy-driven contraction of the stretched dendrimer branches bringing the polymer back to the pore entrance. When the potential was reversed to −100 mV at the trans side, no large current fluctuations were observed. This is in line with the model on the voltage-triggered movement of the dendrimer. Accordingly, the positive PAMAM ball would be expected to move toward the negative pole at the trans side and remain at the pore entrance without subsequent movement out of the channel. The voltage-dependent dynamic behavior of the tethered PAMAM dendrimer indicates that it could function as a voltage-sensing molecular-ball valve.

Current traces also highlighted the different structural dynamics of compact PAMAM as opposed to flexible PEG polymers at the single-molecule level. Pores with a single G2 or G3 PAMAM inside the lumen had a constant current implying that conformational changes of the dendrimer were too small to alter the conductance. In contrast, pores modified with a single PEG chain (5 kD) of comparable hydrodynamic volume to G3

(34) Kasianowicz, J. J. *Nat. Mater.* **2004**, *3*, 355−356. Akeson, M.; Branton, D.; Kasianowicz, J. J.; Brandin, E.; Deamer, D. W. *Biophys. J.* **1999**, *77*, 3227−3233.

PAMAM exhibited short large amplitude events suggesting significant changes in the polymer structure. The recordings, which were performed in another study,[24] displayed events characterized by a current blockade of 51 ± 3%, an average duration of 14 ± 2 ms, and a frequency of occurrence of 0.2 ± 0.02 s^{-1}. The short events originated from the threading of the free end of the polymer chain into a narrow constriction of the pore through the β-barrel toward the trans side of the pore (Figure 1B).[24,25,31] Our single-molecule data therefore offer a glimpse of the different structural dynamics of individual "hard" positively charged PAMAM dendrimers as compared to "soft" neutral PEG chains.

Discussion

This report describes the generation of sulfhydryl-reactive PAMAM dendrimers and their coupling to cysteine mutants of a protein pore of known structure. The extent of reaction of the dendrimers was found to depend on the PAMAM generation and position of the engineered residues in the protein pore. G5 with a hydrodynamic diameter of 6.2 nm did not enter the cis entrance, while G3 with 4.2 nm coupled inside the pore. This clear cutoff in the permeation properties is in contrast to the greater accessibility of flexible PEG chains. For example, PEG-MAL 10 and 5 kD with comparable hydrodynamic diameters of 6.2 and 4.4 nm[35] coupled inside the pore lumen. The results of the present study are in line with another report on the use of PEG. Using the same αHL pore as a model system it was found that linear PEG polymers 1−3 kD (hydrodynamic diameters of 2.1−3.0 nm) could pass through the 1.3 nm wide inner constriction of αHL.[27] Only PEG 5 kD with a diameter of 4.4 nm translocated poorly. Hence, a PEG chain was only restricted in its permeation when the hydrodynamic parameter was at least 3.5 times higher than the pore constriction. By contrast, this value is 2 for G5-PAMAM and expected to be even lower for higher generation dendrimers with more rigid structure.

The sulfhydryl-reactive dendrimers presented in this study are a new type of research reagent for examination of the molecular structure of proteins. The reagents can be used in the substituted-cysteine accessibility method (SCAM),[26] which infers the surface accessibility of residues by determining how fast sulfhydryl-reactive reagents couple to single-cysteine mutants of a protein. SCAM has been widely exploited to probe the structure of membrane proteins and ion channels in combination with patch clamp or lipid bilayer recordings,[26,36] and gel electrophoretic mobility shift assays.[24,27,37,38] Despite the wide range of sulfhydryl-active organic reagents or polymeric reagents, most are too small or too flexible[27] for membrane proteins with pore diameters of more than 2−3 nm because they can either access all pore-lining residues or their coupling to cysteine residues does not give rise to an appreciable current block or gel shift, making it difficult to detect successful modification. Sulfhydryl-reactive PAMAM dendrimers with diameters of several nanometers can overcome these constraints as shown in this study. In calibration experiments with the structurally defined membrane pore αHL, dendrimers exhibited a sharp size-dependent permeation cutoff and readily identified a 2.9 nm wide pore entrance. PAMAM-PDP dendrimers with diameters from 2 to 10 nm are well suited to probe various nanometer-sized pore-forming proteins with important biomedical roles such as bacterial pore-forming toxins,[39] pores of the complement system, or dilating purinergic receptors.[40] In addition, the approach can also be applied to explore the molecular structure of interesting biological nanomaterials such as porous S-layer proteins.[38,41]

The study presents a new way to alter the properties of proteins via targeted chemical modification with hyperbranched dendrimers. Coupling of nonbranched linear organic polymers or biopolymers has been used in the past to modify or expand the natural characteristics of proteins such as in pharmacology to increase the half-life of therapeutic proteins via PEGylation,[42] in molecular biology to introduce sequence specificity into nucleases via attachment of a DNA oligonucleotides,[43] or in microarray technology to achieve targeted immobilization of PNA-modified proteins onto DNA-microarrays.[44] While dendrimers have been attached to proteins to produce new types of immunoreagents with enhanced binding capacity,[45] their use to alter the conductance properties of pores is new. Our results show that a dendrimer molecule filled at least the lumen of the engineered pore and thus regulated the ion flux.

The hyperbranched dendrimers were demonstrated to function as ion-selectivity filters for the passage of small ions through the pore. As expected, the positively charged primary amines on the surface of the PAMAM dendrimer led to a preferred permeation of anions over cations. The advantage of PAMAM over other ion-selective elements such as cyclodextrins[32] is that the dendrimer filters are available in different sizes up to 10 nm. This offers the possibility to use dendrimers for inorganic pores with wide lumen. PAMAM was also demonstrated to function as a molecular sieve to control the passage of molecules through the pore based on their molecular weight. Our results show that the RNA oligonucleotides were blocked while smaller ions were only minimally affected. It is expected that the molecular weight cut off of the molecular sieves can be tuned by increasing or decreasing the size of the dendrimer or changing its chemical composition. While movement of RNA through G3 PAMAM-filled pores is blocked, use of smaller dendrimers—in combination with higher potentials—could possibly enable passage of nucleic acids at reduced translocation speeds with applications in DNA sensing.[46]

Our approach to alter pore permeation properties by placing a spherical dendrimer inside the lumen is new. Use of hyperbranched dendrimers has specific advantages over other spheri-

(35) Scherrer, R.; Gerhardt, P. *J. Bacteriol.* **1971**, *107*, 718−735.
(36) Akabas, M. H.; Stauffer, D. A.; Xu, M.; Karlin, A. *Science* **1992**, *258*, 307−310.
(37) Walker, B.; Bayley, H. *J. Biol. Chem.* **1995**, *270*, 23065−23071. Lu, J.; Deutsch, C. *Biochemistry* **2001**, *40*, 13288−13301.
(38) Howorka, S.; Sára, M.; Wang, Y.; Kuen, B.; Sleytr, U. B.; Lubitz, W.; Bayley, H. *J. Biol. Chem.* **2000**, *48*, 37876−37886.
(39) Parker, M. W.; Feil, S. C. *Prog. Biophys. Mol. Biol.* **2005**, *88*, 91−142. Menestrina, G. *Pore-forming peptides and protein toxins*; Taylor & Francis, CRC: Boca Raton, FL, 2003.
(40) Surprenant, A.; Rassendren, F.; Kawashima, E.; North, R. A.; Buell, G. *Science* **1996**, *272*, 735−738. Khakh, B. S.; North, R. A. *Nature* **2006**, *442*, 527−532.
(41) Sleytr, U. B.; Messner, P.; Pum, D.; Sara, M. *Angew. Chem., Int. Ed.* **1999**, *38*, 1035−1054.
(42) Veronese, F. M.; Pasut, G. *Drug. Discovery Today* **2005**, *10*, 1451−1458.
(43) Corey, D. R.; Schultz, P. G. *Science* **1987**, *238*, 1401−1403.
(44) Winssinger, N.; Harris, J. L.; Backes, B. J.; Schultz, P. G. *Angew. Chem., Int. Ed. Engl.* **2001**, *40*, 3152−3155.
(45) Singh, P. *Bioconjugate Chem.* **1998**, *9*, 54−63. Patri, A. K.; Myc, A.; Beals, J.; Thomas, T. P.; Bander, N. H.; Baker, J. R., Jr. *Bioconjugate Chem.* **2004**, *15*, 1174−1181. Kobayashi, H.; Kawamoto, S.; Saga, T.; Sato, N.; Ishimori, T.; Konishi, J.; Ono, K.; Togashi, K.; Brechbiel, M. W. *Bioconjugate Chem.* **2001**, *12*, 587−593.
(46) Bayley, H. *Curr. Opin. Biotechnol.* **2006**, *10*, 628−637.

cal materials such as quantum dots of similar size or functionality. First, it would be difficult to couple a solid quantum dot into a pore lumen of similar diameter because the hard, highly symmetrical sphere would likely clash with the corrugated protein surface of the less symmetrical, imperfectly circular pore opening. By contrast, compact dendrimers have some residual degree of structural flexibility which can help overcome small steric permeation barriers. Indeed, G3 PAMAM with a diameter of 2.9 nm successfully passed the 2.9 nm wide opening of the αHL pore. Furthermore, the hyperbranched character of the dendrimer is important in controlling and tuning the flow of matter through the engineered pore, while solid impermeable spheres would likely lead to much more drastic and less tunable changes in the permeation properties. In summary, placing dendrimers into a pore lumen is a unique approach to introduce ion-selectivity filters or molecular sieves. The approach is not only restricted to protein pores but might be applied to engineer the permeation properties of inorganic or metallic porous structures for the separation of biopolymers or linear polymers for purification or sensing purposes.

Acknowledgment. We thank Hagan Bayley for providing equipment for lipid bilayer recordings and commenting on the manuscript. H.K. holds a Ph.D. studentship from the U.K. Biotechnology and Biological Sciences Research Council. This work has been supported by the Nuffield Foundation and the Department of Chemistry, UCL.

Supporting Information Available: Experimental procedures, HPLC and MALDI-MS analysis of SPDP-modified dendrimers of generations 2 and 5, quantification of the yield of the PAMAM-PDP preparations, coupling of G3-PAMAM-PDP and G5-PAMAM-PDP to homoheptamer $K8C_7$, coupling of MAL-PEG 10 kD to homoheptamer $S106C_7$, and $I-V$ curves for $S106C_7$, $S106C_7$-G2-PAMAM, and $S106C_7$-G3-PAMAM. This material is available free of charge via the Internet at http://pubs.acs.org.

JA0689029

Appendix B

Image Work

B.1 Published Images

This section lists the images I have produced that featured in journal publications.

Nanopore analytics: sensing of single molecules, Howorka, S. and Siwy, Z., Chemical Society Reviews, Figure 2. [132].

Handbook of single-molecule biophysics, Hinterdorfer, P. and Van Oijen, A., Figure 11.2 [133].

Stochastic Detection of Motor Protein-RNA Complexes by Single-Channel Current Recording, Chem Phys Chem, Astier, Y., Kainov, D.E., Bayley, H., Tuma, R. and Howorka, S., Figure 1 [134]

Chemical Tags Facilitate the Sensing of Individual DNA Strands with Nanopores, Mitchell, N. and Howorka, S., Angewandte Chemie, Figure. 1A [135]

Chemical tags mediate the orthogonal self-assembly of DNA strands into supramolecular structures. Mitchell, N., Ebner, A., Hinterdorfer, P., Tamp, R., Howorka, S., Small, Figure 7B and 7C Supporting information [136]

A DNA Nanostructure for the Functional Assembly of Chemical Groups with Tunable Stoichiometry and Defined Nanoscale Geometry, Mitchell, N., Schlapak, R., Kastner, M., Armitage, D., Chrzanowski, W., Riener, J., Hinterdorfer, P., Ebner, A., Howorka, S., Angewandte Chemie, Figure 1A and 1C [137]

B.2 Cover Images

This section shows the images I have produced that featured on the covers of journals.

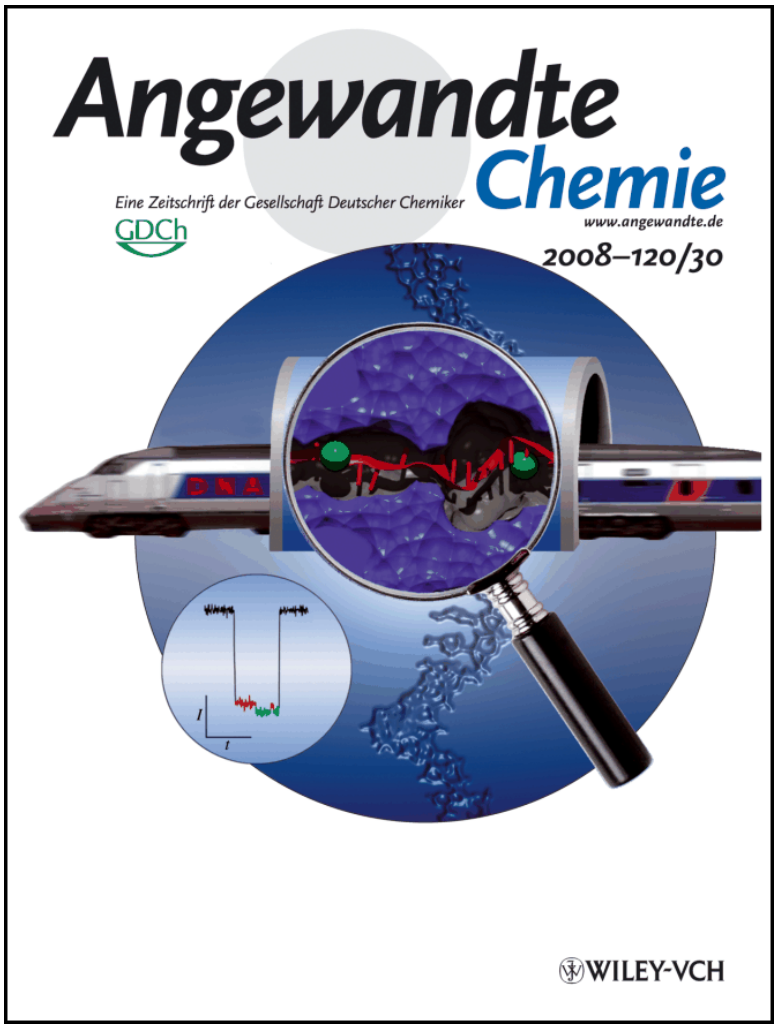


FIGURE B.1: Inside Cover: Chemical Tags Facilitate the Sensing of Individual DNA Strands with Nanopores, Mitchell, N. and Howorka, S., Angewandte Chemie [138]. Image prepared using VMD [3].

FIGURE B.2: Cover: Selective protein and DNA adsorption on PLL-PEG films modulated by ionic strength, Schlapak, R., Armitage, D., Saucedo-Zeni, N., Chrzanowski, W., Hohage, M., Caruana, D. and Howorka, S., Soft Matter [139]. Image prepared using VMD [3].

FIGURE B.3: Cover: Determination of Free Energy Profiles for the Translocation of Polynucleotides through α-Hemolysin Nanopores using Non-Equilibrium Molecular Dynamics Simulations, Martin, H.S.C. and Jha, S. and Howorka, S. and Coveney, P.V., JCTC [125]. Image prepared using VMD [3] and Blender [140].

B.3 Featured Images

This section shows the images I have produced that featured outside of publications.

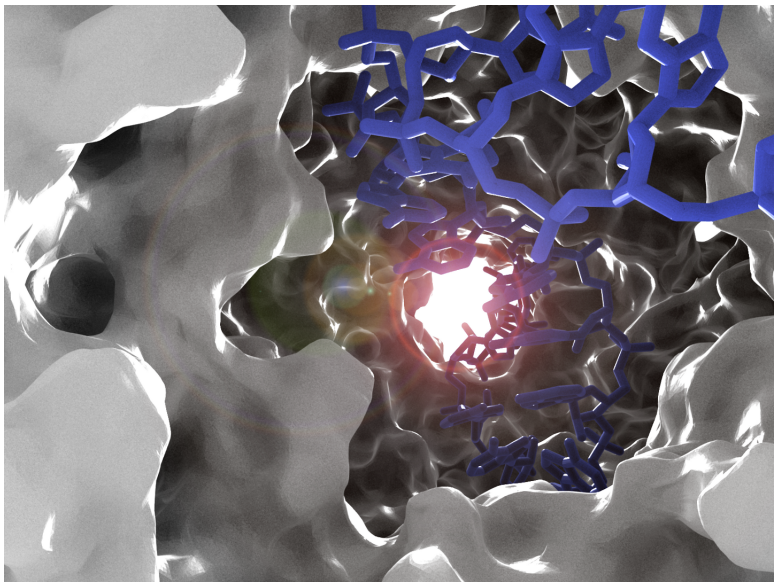


FIGURE B.4: UCL Research Images as Art Competition 2009 [141]. An internal view of the alpha-hemolysin protein-pore (grey). A 25-base adenine polynucleotide molecule (blue) translocates through the pore during a constant velocity-steered molecular dynamics simulation. The atomic positions used to create this image are taken directly from a simulation. Image prepared using VMD [3] and Blender [140].

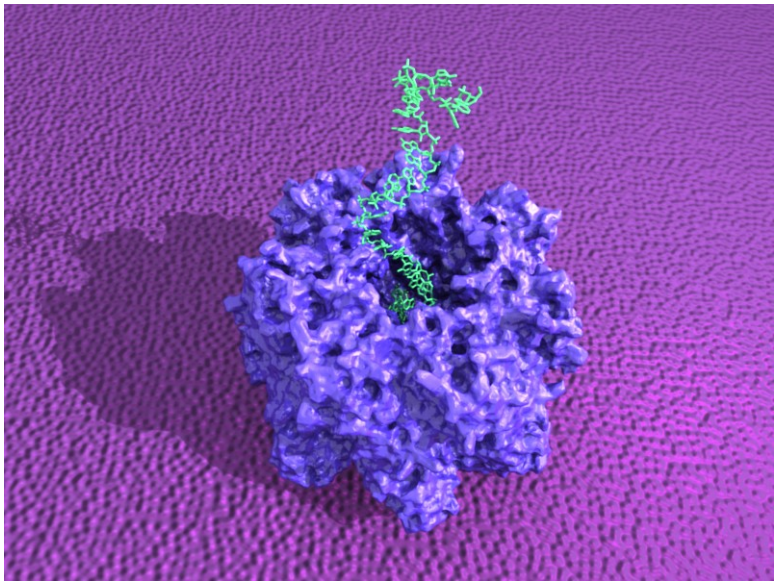


FIGURE B.5: CCS website image [142]. A polyadenosine molecule (green) translocates through an αHL protein pore (blue), which is inserted into a lipid membrane (purple). Image prepared using VMD [3] and Blender [140].

Appendix C

Alpha-Hemolysin and P4 ATPase

This chapter outlines an unfinished project, started in 2007 during my PhD. The aim of the project was to couple P4 ATPase of bacteriophage $\phi 8$ to the *cis*-side of αHL to generate a nanopore complex that can control the speed of polynucleotide translocation. The ability to regulate and slow down the translocation is relevant to DNA sequencing using nanopore current recording.

Currently, a major problem in nanopore sequencing is the fast velocity at which the polymers translocate through pores such as αHL, leading to insufficient resolution in current traces. For example, nucleic acid chains composed of 100 bases pass through the pore within a couple of milliseconds [49]. The signal to noise ratio at this speed, while allowing the clear discrimination of pore blockage, does not allow for the distinction of single bases in a polymer chain. One approach to slow down the polynucleotide translocation is the use of proteins that bind nucleic acids and thereby act as molecular breaks. This idea was put forward by Astier *et al.* [134] and Hornblower *et al.* [143]. The experiments by Astier *et al.* demonstrated that the RNA binding protein P4 ATPase can slow down polynucleotide translocation through αHL via the formation of a transient complex, which is shown in Figure C.1. In this complex, the P4 protein holds on to the RNA strand thereby retarding the passage of the polynucleotide.

The viral package motor P4 of bacteriophage $\phi 8$ is a hexameric protein that forms a pore through which RNA is translocated in an ATP dependant mechanism [144, 145]. The $\phi 8$ form of P4 is particularly useful in this instance due to its relatively strong RNA

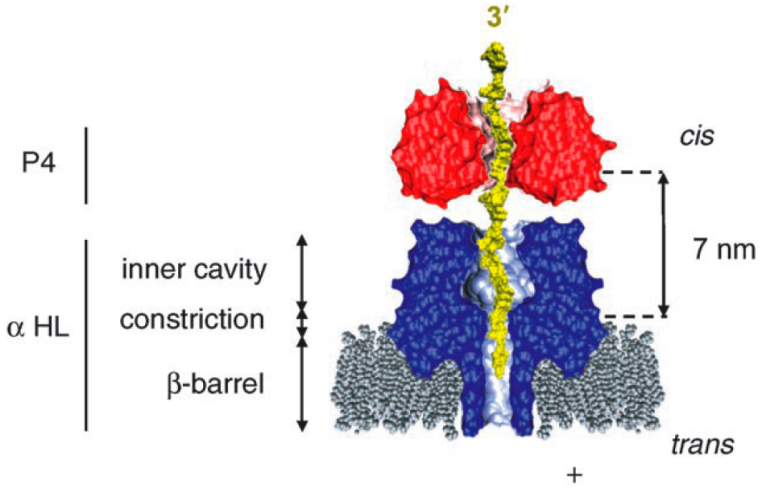


FIGURE C.1: Proposed interaction of the viral RNA translocase P4 with αHL. A cross section of the heptameric membrane protein pore aHL inserted into a lipid bilayer. The model of the dome-shaped P4 translocase is based on the crystal structure of P4 from bacteriophage ϕ12, which has a high extent of sequence similarity to P4 from ϕ8. The yellow strand indicates a polynucleotide strand threading through the αHL-P4 complex [134].

binding, translocation, and complementary oligonucleotide displacement activity [146–148]. P4 is not a transmembrane protein, therefore, in order to be used in single channel current recording experiments, it must be used in conjunction with a transmembrane nanopore such as αHL.

The aim of the αHL-P4 project in my PhD was to generate a chemically coupled complex between P4 and αHL. The coupling was to be mediated via a metal chelate bridge between genetically engineered hexahistidine affinity tags at P4 and complementary NTA tags that are chemically attached to cysteine residues at the *cis*-side of the αHL pore. It was envisioned that the resulting nanopore complex would retain the ability of P4 to slow down the translocation. Unfortunately, experimental problems prevented the project from being successfully pursued. Firstly, the attempts to couple P4 to αHL via the NTA groups indicated a degree of non-specific binding. This undesired feature could not be avoided by more stringent washing conditions because these led to the disintegration of the weakly bound complex. Secondly, the P4 protein was prone to proteolysis under conditions used to form the complex on erythrocyte membranes. With sufficient time, it may have been possible to overcome these issues. However, the demands of the computational aspects of this thesis were too great, and so the αHL-P4 project was abandoned.

Bibliography

[1] J. Šponer and F. Lankaš. *Computational Studies of RNA and DNA*. Springer Verlag, 2006.

[2] J. Šponer, J. Leszczynski, and P. Hobza. Electronic properties, hydrogen bonding, stacking, and cation binding of dna and rna bases. *Biopolymers*, 61(1):3–31, 2001. ISSN 1097-0282.

[3] W. Humphrey, A. Dalke, and K. Schulten. VMD-Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[4] A. Adler, L. Grossman, and GD Fasman. Single-stranded oligomers and polymers of cytidylic and 2'-deoxycytidylic acids: Comparative optical rotatory studies. *Proceedings of the National Academy of Sciences*, 57(2):423–430, 1967.

[5] DN Holcomb and I. Tinoco Jr. Conformation of polyriboadenylic acid: pH and temperature dependence. *Peptide Science*, 3(2):121–133, 1965.

[6] J. Brahms, AM Michelson, and KE Van Holde. Adenylate oligomers in single-and double-strand conformation. *Journal of Molecular Biology*, 15(2):467–88, 1966.

[7] M. Leng and G. Felsenfeld. A study of polyadenylic acid at neutral ph. *Journal of Molecular Biology*, 15(2):455–66, 1966.

[8] E. Buxbaum. *Fundamentals of protein structure and function*. Springer Verlag, 2007. ISBN 0387263527.

[9] Yada, R.Y. and Jackman, R.L. and Nakai, S. Secondary structure prediction and determination of proteins-a review. *International Journal of Peptide and Protein Research*, 31(1):98–108, 1988.

[10] B. Dreiseikelmann. Translocation of dna across bacterial membranes. *Microbiology and Molecular Biology Reviews*, 58(3):293–316, 1994.

[11] H. Nikaido. Molecular basis of bacterial outer membrane permeability revisited. *Microbiology and Molecular Biology Reviews*, 67(4):593–656, 2003.

[12] B. Hanss, E. Leal-Pinto, L.A. Bruggeman, T.D. Copeland, and P.E. Klotman. Identification and characterization of a cell membrane nucleic acid channel. *Proceedings of the National Academy of Sciences*, 95(4):1921–1926, 1998.

[13] Tzfira, T. and Rhee, Y. and Chen, M.H. and Kunik, T. and Citovsky, V. Nucleic acid transport in plant-microbe interactions: The molecules that walk through the walls. *Annual Review of Microbiology*, 2003.

[14] V. Citovsky and P. Zambryski. Transport of nucleic acids through membrane channels: snaking through small holes. *Annual Reviews in Microbiology*, 47(1): 167–197, 1993.

[15] G. Whittaker. Squeezing through pores: how microorganisms get into and out of the nucleus. *Trends in Microbiology*, 6(5):178–179, 1998.

[16] F.S. Collins, E.D. Green, A.E. Guttmacher, M.S. Guyer, et al. A vision for the future of genomics research. *Nature*, 422(6934):835–847, 2003.

[17] F.S. Collins. Medical and Societal Consequences of the Human Genome Project, 1999.

[18] F.S. Collins and V.A. McKusick. Implications of the Human Genome Project for Medical Science, 2001.

[19] G.J.B. van Ommen. The Human Genome Project and the future of diagnostics, treatment and prevention. *Journal of Inherited Metabolic Disease*, 25(3):183–188, 2002.

[20] A.E. Guttmacher and F.S. Collins. Genomic Medicine-A Primer. *The New England Journal of Medicine*, 347(19):1512–1520, 2002.

[21] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447, June 2007.

[22] M.R. Stratton, P.J. Campbell, and P.A. Futreal. The cancer genome. *Nature*, 458 (7239):719–724, 2009.

[23] GN Godson. Sequencing DNA by the Sanger chain termination method. *Methods of DNA and RNA sequencing. Praeger Publishers, New York*, pages 69–111, 1983.

[24] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[25] Genomes OnLine Database, 2010. URL http://www.genomesonline.org.

[26] G.M. Church. Genomes for ALL. *Scientific American*, 294(1):46–54, 2006.

[27] R. Drmanac, A.B. Sparks, M.J. Callow, A.L. Halpern, N.L. Burns, B.G. Kermani, P. Carnevali, I. Nazarenko, G.B. Nilsen, G. Yeung, et al. Human genome sequencing using unchained base reads on self-assembling dna nanoarrays. *Science*, 327 (5961):78, 2010.

[28] National Institutes of Health, 2006. URL http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-05-004.html.

[29] Reuters Special Report, 2010. URL http://www.reuters.com/article/idUSTRE62T0KC20100330.

[30] ACS, 2009. URL http://pubs.acs.org/cen/coverstory/87/8750cover2.html.

[31] I. Braslavsky, B. Hebert, E. Kartalov, and S.R. Quake. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences*, 100(7):3960–3964, 2003.

[32] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, et al. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133, 2009.

[33] H. Bayley. Sequencing single molecules of dna. *Current Opinion in Chemical Biology*, 10(6):628–637, 2006.

[34] J. Clarke, H.C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley. Continuous base identification for single-molecule nanopore dna sequencing. *Nature Nanotechnology*, 4:265–270, 2009.

[35] SB Hladky and DA Haydon. Discreteness of conductance change in bimolecular lipid membranes in the presence of certain antibiotics. *Nature*, 1970.

[36] E. Neher and B. Sakmann. Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature*, 260(5554):799–802, 1976.

[37] J.J. Kasianowicz, E. Brandin, D. Branton, and D.W. Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24):13770–13773, 1996.

[38] L. Song, M.R. Hobaugh, C. Shustak, S. Cheley, H. Bayley, and J.E. Gouaux. Structure of Staphylococcal alpha-Hemolysin, a Heptameric Transmembrane Pore. *Science*, 274(5294):1859, 1996.

[39] W.L. DeLano. The PyMOL Molecular Graphics System. DeLano Scientific, 2002.

[40] V. Borisenko, T. Lougheed, J. Hesse, E. Füreder-Kitzmüller, N. Fertig, JC Behrends, GA Woolley, and GJ Schütz. Simultaneous optical and electrical recording of single gramicidin channels. *Biophysical Journal*, 84(1):612–622, 2003.

[41] M. Mayer, V. Semetey, I. Gitlin, J. Yang, and G.M. Whitesides. Using ion channel-forming peptides to quantify protein- ligand interactions. *J. Am. Chem. Soc*, 130 (4):1453–1465, 2008.

[42] M. Chen, S. Khalid, M.S.P. Sansom, and H. Bayley. Outer membrane protein g: Engineering a quiet pore for biosensing. *Proceedings of the National Academy of Sciences*, 105(17):6272, 2008.

[43] T.Z. Butler, M. Pavlenok, I.M. Derrington, M. Niederweis, and J.H. Gundlach. Single-molecule dna detection with an engineered mspa protein nanopore. *Proceedings of the National Academy of Sciences*, 105(52):20647, 2008.

[44] C.A. Merchant, K. Healy, M. Wanunu, V. Ray, N. Peterman, J. Bartel, M.D. Fischbein, K. Venta, Z. Luo, A.T.C. Johnson, et al. Dna translocation through graphene nanopores. *Nano Letters*, pages 580–1132, 2010.

[45] J. Li, D. Stein, C. McMullan, D. Branton, M.J. Aziz, and J.A. Golovchenko. Ion-beam sculpting at nanometre length scales. *Nature*, 412(6843):166–169, 2001.

[46] AJ Storm, JH Chen, XS Ling, HW Zandbergen, and C. Dekker. Fabrication of solid-state nanopores with single-nanometre precision. *Nature materials*, 2(8): 537–540, 2003.

[47] J.B. Heng, C. Ho, T. Kim, R. Timp, A. Aksimentiev, Y.V. Grinkova, S. Sligar, K. Schulten, and G. Timp. Sizing DNA Using a Nanometer-Diameter Pore. *Biophysical Journal*, 87(4):2905–2911, 2004.

[48] S.M. Iqbal, D. Akin, and R. Bashir. Solid-state nanopore channels with dna selectivity. *Nature Nanotechnology*, 2(4):243–248, 2007.

[49] M. Akeson, D. Branton, J.J. Kasianowicz, E. Brandin, and D.W. Deamer. Microsecond Time-Scale Discrimination Among Polycytidylic Acid, Polyadenylic Acid, and Polyuridylic Acid as Homopolymers or as Segments Within Single RNA Molecules. *Biophysical Journal*, 77(6):3227–3233, 1999.

[50] A. Meller, L. Nivon, E. Brandin, J. Golovchenko, and D. Branton. Rapid nanopore discrimination between single polynucleotide molecules. *Proceedings of the National Academy of Sciences*, 97(3):1079–1084, 2000.

[51] T.Z. Butler, J.H. Gundlach, and M.A. Troll. Determination of RNA Orientation during Translocation through a Biological Nanopore. *Biophysical Journal*, 90(1): 190–199, 2006.

[52] A. Meller, L. Nivon, and D. Branton. Voltage-Driven DNA Translocations through a Nanopore. *Physical Review Letters*, 86(15):3435–3438, 2001.

[53] J. Mathe, H. Visram, V. Viasnoff, Y. Rabin, and A. Meller. Nanopore Unzipping of Individual DNA Hairpin Molecules. *Biophysical Journal*, 87(5):3205–3212, 2004.

[54] S. Howorka, S. Cheley, and H. Bayley. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nature Biotechnology*, 19:636–639, 2001.

[55] Y. Astier, O. Braha, and H. Bayley. Toward Single Molecule DNA Sequencing: Direct Identification of Ribonucleoside and Deoxyribonucleoside 5'-Monophosphates by Using an Engineered Protein Nanopore Equipped with a Molecular Adapter. *Science*, 309:1728–1732, 2005.

[56] J. Mathe, A. Aksimentiev, D.R. Nelson, K. Schulten, and A. Meller. Orientation discrimination of single-stranded DNA inside the α-hemolysin membrane channel. *Proceedings of the National Academy of Sciences*, 102(35):12377–12382, 2005.

[57] H. Wang, J.E. Dunning, A.P.H. Huang, J.A. Nyamwanda, and D. Branton. DNA heterogeneity and phosphorylation unveiled by single-molecule electrophoresis. *Proceedings of the National Academy of Sciences*, 101(37):13472–13477, 2004.

[58] N. Ashkenasy, J. Sánchez-Quesada, MR Ghadiri, and H. Bayley. Recognizing a Single Base in an Individual DNA Strand: A Step Toward Nanopore DNA Sequencing. *Angewandte Chemie-International Edition*, 44(9):1401–1404, 2005.

[59] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 2002.

[60] P.L. Freddolino, F. Liu, M. Gruebele, and K. Schulten. Ten-microsecond molecular dynamics simulation of a fast-folding ww domain. *Biophysical Journal*, 94(10):75–77, 2008.

[61] P. Maragakis, K. Lindorff-Larsen, M.P. Eastwood, R.O. Dror, J.L. Klepeis, I.T. Arkin, M.O. Jensen, H. Xu, N. Trbovic, R.A. Friesner, et al. Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins. *Journal of Physical Chemistry B*, 112(19):6155, 2008.

[62] D.E. Shaw, R.O. Dror, J.K. Salmon, JP Grossman, K.M. Mackenzie, J.A. Bank, C. Young, M.M. Deneroff, B. Batson, K.J. Bowers, et al. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, pages 1–11. ACM, 2009.

[63] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y. Shan, et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341, 2010. ISSN 0036-8075.

[64] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[65] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput*, 4(3):435–447, 2008.

[66] K.J. Bowers, E. Chow, H. Xu, R.O. Dror, M.P. Eastwood, B.A. Gregersen, J.L. Klepeis, I. Kolossvary, M.A. Moraes, F.D. Sacerdoti, et al. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, page 84. ACM, 2006.

[67] S.E. Feller and AD MacKerell. An improved empirical potential energy function for molecular simulations of phospholipids. *The Journal of Physical Chemistry B*, 104(31):7510–7515, 2000.

[68] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham III, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3):1–41, 1995.

[69] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3): 712–725, 2006. ISSN 1097-0134.

[70] R.W. Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *Journal of Chemical Physics*, 22:1420–1426, 1954.

[71] C. Chipot and A. Pohorille. *Free energy calculations: theory and applications in chemistry and biology*. Springer Berlin, 2007.

[72] J.G. Kirkwood. *Theory of liquids*. Gordon and Breach New York, 1968.

[73] J.G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3:300, 1935.

[74] G.M. Torrie and J.P. Valleau. Monte carlo free energy estimates using non-boltzmann sampling: Application to the sub-critical lennard-jones fluid. *Chemical Physics Letters*, 28:578–581, 1974.

[75] GM Torrie and JP Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: umbrella sampling. *Journal of Computational Physics*, 23 (2):187–199, 1977.

[76] E. Darve and A. Pohorille. Calculating free energies using average force. *The Journal of Chemical Physics*, 115:9169, 2001.

[77] E. Darve, M.A. Wilson, and A. Pohorille. Calculating free energies using a scaled-force molecular dynamics algorithm. *Molecular Simulation*, 28(1-2):113–144, 2002.

[78] J. Hénin and C. Chipot. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *The Journal of Chemical Physics*, 121:2904, 2004.

[79] C. Chipot and J. Hénin. Exploring the free-energy landscape of a short peptide using an average force. *The Journal of Chemical Physics*, 123:244906, 2005.

[80] J. Hénin, G. Fiorin, C. Chipot, and M.L Klein. Exploring Multidimensional Free Energy Landscapes Using Time-Dependent Biases on Collective Variables. *Journal of Chemical Theory and Computation*, 6(1):164–170, 2010.

[81] J. Henin, E. Tajkhorshid, K. Schulten, and C. Chipot. Diffusion of glycerol through Escherichia coli aquaglyceroporin GlpF. *Biophysical Journal*, 94(3):832–839, 2008.

[82] M.O. Jensen, S. Park, E. Tajkhorshid, and K. Schulten. Energetics of glycerol conduction through aquaglyceroporin GlpF. *Proceedings of the National Academy of Sciences*, 99(10):6731, 2002.

[83] W. Treptow and M. Tarek. Molecular restraints in the permeation pathway of ion channels. *Biophysical Journal*, 91(3):L26–L28, 2006.

[84] I. Ivanov, X. Cheng, S.M. Sine, J.A. McCammon, et al. Barriers to ion translocation in cationic and anionic receptors from the cys-loop family. *Journal of the American Chemical Society*, 129(26):8217–8224, 2007.

[85] G. Lamoureux, M.L. Klein, and S. Bernèche. A stable water chain in the hydrophobic pore of the amtb ammonium transporter. *Biophysical Journal*, 92(9): 82–84, 2007.

[86] C. Wei and A. Pohorille. Permeation of Membranes by Ribose and Its Diastereomers. *Journal of the American Chemical Society*, 131(29):10237, 2009.

[87] J. Xu, M.F. Crowley, and J.C. Smith. Building a foundation for structure-based cellulosome design for cellulosic ethanol: Insight into cohesin-dockerin complexation from computer simulation. *Protein Science*, 18(5):949–959, 2009.

[88] W. Cai, T. Sun, P. Liu, C. Chipot, and X. Shao. Inclusion Mechanism of Steroid Drugs into β-Cyclodextrins. Insights from Free Energy Calculations. *J. Phys. Chem. B*, 113(22):7836–7843, 2009.

[89] J. Rodriguez, R. Semino, and D. Laria. Building up Nanotubes: Docking of" Janus" Cyclodextrins in Solution. *The Journal of Physical Chemistry. B*, 113(5): 1241, 2009.

[90] S. Vaitheeswaran and D. Thirumalai. Interactions between amino acid side chains in cylindrical hydrophobic nanopores with applications to peptide stability. *Proceedings of the National Academy of Sciences*, 105(46):17636, 2008.

[91] F. Dehez, E. Pebay-Peyroula, and C. Chipot. Binding of ADP in the mitochondrial ADP/ATP carrier is driven by an electrostatic funnel. *Journal of the American Chemical Society*, 130(38):12725–12733, 2008.

[92] L. Kale, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten. NAMD2: Greater scalability for parallel molecular dynamics. *Journal of Computational Physics*, 151(1):283–312, 1999.

[93] C. Jarzynski. Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters*, 78(14):2690–2693, 1997.

[94] S. Park and F. Khalili-Araghi. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *The Journal of Chemical Physics*, 119(6):3559, 2003.

[95] S. Park. Calculating potentials of mean force from steered molecular dynamics simulations. *The Journal of Chemical Physics*, 120(13):5946, 2004.

[96] S. Mukamel. Quantum Extension of the Jarzynski Relation: Analogy with Stochastic Dephasing. *Physical Review Letters*, 90(17):170604, 2003.

[97] G. Hummer and A. Szabo. From the Cover: Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proceedings of the National Academy of Sciences*, 98(7):3658, 2001.

[98] F. Douarche, S. Ciliberto, A. Petrosyan, and I. Rabbiosi. An experimental test of the Jarzynski equality in a mechanical experiment. *Europhys. Lett*, 70(5):593–599, 2005.

[99] G.E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3):2721–2726, 1999.

[100] D.J. Evans, E.G.D. Cohen, and G.P. Morriss. Probability of second law violations in shearing steady states. *Physical Review Letters*, 71(15):2401–2404, 1993.

[101] G.D. Fabritiis, P.V. Coveney, and J. Villa-Freixa. Energetics of K^+ permeability through Gramicidin A by forward-reverse steered molecular dynamics. *Proteins*, 73:185–194, 2008.

[102] M. Karplus and J.A. McCammon. Molecular Dynamics. *Nature Structural Biology*, 9:646–652, 2002.

[103] D. Rognan. Molecular dynamics simulations: a tool for drug design. *Perspectives in Drug Discovery and Design*, 9:181–209, 1998.

[104] M. Karplus and J.A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652, 2002.

[105] A. Aksimentiev and K. Schulten. Imaging α-Hemolysin with Molecular Dynamics: Ionic Conductance, Osmotic Permeability, and the Electrostatic Potential Map. *Biophysical Journal*, 88(6):3745–3761, 2005.

[106] M. Muthukumar and CY Kong. Simulation of polymer translocation through protein channels. *Proceedings of the National Academy of Sciences*, 103(14):5273–5278, 2006.

[107] D.B. Wells, A. Abramkina, and Aksimentiev A. Exploring transmembrane transport through α-hemolysin with grid-steered molecular dynamics. *The Journal of Chemical Physics*, 127:125101–1–125101–10, 2007.

[108] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79:926, 1983.

[109] EA Merritt and M.E.P. Murphy. Raster3d version 2.0. a program for photorealistic molecular graphics. *Acta Crystallographica Section D: Biological Crystallography*, 50(6):869–873, 1994. ISSN 0907-4449.

[110] S. Arnott, PJC Smith, and R. Chandrasekharan. *CRC Handbook of Biochemistry and Molecular Biology: Nucleic Acids*. CRC Press, Cleveland, OH, 1972.

[111] M. Wanunu, B. Chakrabarti, J. Mathé, D.R. Nelson, and A. Meller. Orientation-dependent interactions of dna with an α-hemolysin channel. *Physical Review E*, 77(3):31904, 2008.

[112] M. Wiggin, C. Tropini, V. Tabard-Cossa, N.N. Jetha, and A. Marziali. Nonexponential kinetics of dna escape from {alpha}-hemolysin nanopores. *Biophysical Journal*, 95(11):5317, 2008.

[113] P.F. Batcho. Optimized particle-mesh Ewald/multiple-time step integration for molecular dynamics simulations. *The Journal of Chemical Physics*, 115(9):4003, 2001.

[114] A.T. Brünger. *X-Plor Version 3.1: A System for X-Ray Crystallography and NMR*. Yale University Press, 1993.

[115] G.J. Martyna, D.J. Tobias, and M.L. Klein. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics*, 101:4177, 1994.

[116] S. Jha, P. Coveney, and M. Harvey. SPICE: Simulated Pore Interactive Computing Environment. *Proceedings of the 2005 ACM/IEEE Conference on Supercomputing*, page 70, 2005.

[117] B. Boghosian, P. Coveney, S. Dong, L. Finn, S. Jha, G. Karniadakis, and N. Karonis. Nektar, spice and vortonics: using federated grids for large scale scientific applications. *Cluster Computing*, 10(3):351–364, 2007.

[118] B. Boghosian, P. Coveney, S. Dong, L. Finn, S. Jha, G. Karniadakis, and N. Karonis. Nektar, SPICE and Vortonics: Using Federated Grids for Large Scale Scientific Applications, Challenges on Large Applications in Distributed Environments. *Institute of Electrical and Electronics Engineers*, pages 32–42, 2006.

[119] Queen Bee, 2010. URL http://www.loni.org/systems/system.php?system=QueenBee.

[120] HPCx, 2010. URL http://www.hpcx.ac.uk/.

[121] NGS, 2010. URL http://www.grid-support.ac.uk/.

[122] Teragrid, 2010. URL http://www.teragrid.org/.

[123] LONI, 2010. URL http://www.loni.org/systems/.

[124] Y. Astier, O. Braha, and H. Bayley. Toward single molecule dna sequencing: Direct identification of ribonucleoside and deoxyribonucleoside 5 '-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. *J. Am. Chem. Soc*, 128(5):1705–1710, 2006.

[125] H.S.C. Martin, S. Jha, S. Howorka, and P.V. Coveney. Determination of free energy profiles for the translocation of polynucleotides through α-hemolysin nanopores using non-equilibrium molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 5(8):2135–2148, 2009.

[126] J.C. Ma and D.A. Dougherty. The cation-pi interaction. *Chemical Reviews*, 97(5): 1303–1324, 1997.

[127] G. Maglia, M. Restrepo, E. Mikhailova, and H. Bayley. Enhanced translocation of single dna molecules through α-hemolysin nanopores by manipulation of internal charge. *Proceedings of the National Academy of Sciences*, 105(50):19720–19725, December 2008.

[128] M. Bhandarkar, A. Bhatele, E. Bohm, R. Brunner, F. Buelens, C. Chipot, A. Dalke, S. Dixit, G. Fiorin, P. Freddolino, et al. Namd user's guide. *Urbana*, 51: 61801, 2009.

[129] H. Martin, H. Kinns, N. Mitchell, Y. Astier, R. Madathil, and S. Howorka. Nanoscale protein pores modified with pamam dendrimers. *J. Am. Chem. Soc*, 129(31):9640–9649, 2007.

[130] S. Howorka, L. Movileanu, X. Lu, M. Magnon, S. Cheley, O. Braha, and H. Bayley. A protein pore with a single polymer chain tethered within the lumen. *J. Am. Chem. Soc*, 122(11):2411–2416, 2000.

[131] L. Movileanu, S. Howorka, O. Braha, and H. Bayley. Detecting protein analytes that modulate transmembrane movement of a polymer chain within a single protein pore. *Nature Biotechnology*, 18(10):1091–1095, 2000.

[132] S. Howorka and Z. Siwy. Cheminform abstract: Nanopore analytics: Sensing of single molecules. *ChemInform*, 40(48), 2009.

[133] P. Hinterdorfer and A. Van Oijen. *Handbook of single-molecule biophysics*. Springer Verlag, 2009.

[134] Y. Astier, D.E. Kainov, H. Bayley, R. Tuma, and S. Howorka. Stochastic detection of motor protein-rna complexes by single-channel current recording. *ChemPhysChem*, 8(15):2189–2194, 2007.

[135] N. Mitchell and S. Howorka. Chemical tags facilitate the sensing of individual dna strands with nanopores. *Angewandte Chemie*, 120(30):5647–5650, 2008.

[136] N. Mitchell, A. Ebner, P. Hinterdorfer, R. Tampé, and S. Howorka. Chemical tags mediate the orthogonal self-assembly of dna duplexes into supramolecular structures. *Small*, 6(16):1732–1735, 2010.

[137] N. Mitchell, R. Schlapak, M. Kastner, D. Armitage, W. Chrzanowski, J. Riener, P. Hinterdorfer, A. Ebner, and S. Howorka. A dna nanostructure for the functional assembly of chemical groups with tunable stoichiometry and defined nanoscale geometry. *Angewandte Chemie*, 121(3):533–535, 2009.

[138] N. Mitchell and S. Howorka. Innentitelbild: Chemical tags facilitate the sensing of individual dna strands with nanopores (angew. chem. 30/2008). *Angewandte Chemie*, 120(30):5558, 2008.

[139] R. Schlapak, D. Armitage, N. Saucedo-Zeni, W. Chrzanowski, M. Hohage, D. Caruana, and S. Howorka. Selective protein and dna adsorption on pll-peg films modulated by ionic strength. *Soft Matter*, 5(3):613–621, 2009.

[140] T. Roosendaal and S. Selleri. *The Official Blender 2.3 guide: free 3D creation suite for modeling, animation, and rendering*. No Starch Press, 2004. ISBN 1593270410.

[141] UCL, 2009. URL http://www.grad.ucl.ac.uk/comp/2009-2010/research/gallery/index.pht?entryID=6.

[142] CCS, 2010. URL http://ccs.chem.ucl.ac.uk/research/ahl.shtml.

[143] B. Hornblower, A. Coombs, R.D. Whitaker, A. Kolomeisky, S.J. Picone, A. Meller, and M. Akeson. Single-molecule analysis of DNA-protein complexes using nanopores. *Nature Methods*, pages 315–317, 2007.

[144] J. Lisal, T.K.T. Lam, D.E. Kainov, M.R. Emmett, A.G. Marshall, and R. Tuma. Functional visualization of viral molecular motor by hydrogen-deuterium exchange reveals transient states. *Nature Structural & Molecular Biology*, 12(5):460–466, 2005.

[145] D.E. Kainov, R. Tuma, and E.J. Mancini. Hexameric molecular motors: P4 packaging ATPase unravels the mechanism. *Cellular and Molecular Life Sciences (CMLS)*, 63(10):1095–1105, 2006.

[146] D.E. Kainov, M. Pirttimaa, R. Tuma, S.J. Butcher, G.J. Thomas, D.H. Bamford, and E.V. Makeyev. RNA Packaging Device of Double-stranded RNA Bacteriophages, Possibly as Simple as Hexamer of P4 Protein*. *Journal of Biological Chemistry*, 278(48):48084–48091, 2003.

[147] J. Lisal, D.E. Kainov, D.H. Bamford, G.J. Thomas Jr, and R. Tuma. Enzymatic Mechanism of RNA Translocation in Double-stranded RNA Bacteriophages. *Journal of Biological Chemistry*, 279(2):1343, 2004.

[148] J. Lisal and R. Tuma. Cooperative Mechanism of RNA Packaging Motor. *Journal of Biological Chemistry*, 280(24):23157, 2005.